UNIT 4

Statistical summaries

Introduction

'Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.'

Attributed to H.G. Wells, an English science fiction author.

This is the first of two units in the module that deal with statistical ideas. It looks first at some of the kinds of question that can be addressed by statistical methods and then at some issues that arise when appropriate data have been collected. It then provides an introduction to some important statistical techniques for finding numerical summaries (for example, calculating an average or a measure of spread) in order to see more clearly some of the useful information provided by the data.

This unit presents statistical techniques in the context of practical investigations. As you will see, in any purposeful statistical investigation, there are four useful stages that can help you to organise your planning of the tasks that need to be carried out. In fact, they make up a statistical version of the general mathematical modelling cycle that you met in Unit 2.

In this unit you will see how to use these four stages in the context of using calculations to reveal patterns in data. Then, in the second unit on statistics (Unit 11), you will look again at the four stages, this time in the context of statistical pictures (charts and graphs).

1 Questions, questions

You only need to glance at a newspaper, a magazine, television or the internet to see that statistical information is all around you. A key aim of this unit is to present statistical ideas as more than simply facts and techniques – statistical thinking is presented as a helpful way of seeing the world quantitatively (as opposed to qualitatively), and could become a valuable tool in your decision-making toolbox.

A quantitative view uses numbers such as measurements and counts. A qualitative view describes what something is like in words, for example, 'small, medium, large'.

Mathematical thinking can also be viewed in this way and, indeed, many of the remarks about statistics in this unit can be equally applied to mathematics in general.

'Statistics' can be used as either a singular or a plural word. Statistics in its plural form is probably more familiar to you: statistics are numerical facts. Statistics in its singular form – allowing the wording 'statistics is' – refers to statistics, like mathematics, as a scientific subject. The former are part of the concern of the latter!

Here are some of the ways in which statistics is unavoidable in our lives.

- *Numbers*: each person operates within a variety of key life roles, such as at work, at home, as a consumer and in the wider community. In each of these environments, you are presented with information, often in the form of numbers, that must be processed and interpreted if you are to be a successfully functioning worker, family member, consumer and citizen.
- *Graphs and charts*: statistical information often takes a visual form. You need to know how to interpret these 'data pictures', both in terms of the overall trends and patterns they suggest and also by knowing how to pull out and examine some of the relevant detail.

In fact, increasingly, almost every subject that you might wish to study has become more quantitative, making it ever more important to have a sound grasp of basic statistics.

Much of this statistical information arises as an attempt to *answer* questions of various kinds. For example, should people stop smoking? Should we drive more carefully? But they often end up *raising* just as many questions as they answer!

1.1 Types of statistical question

Before rushing into answering any question, it is always a good idea to ask: 'Have I seen one like this before?' Most of the questions and forms of investigation that occur in statistics can be categorised into one of the following three types.

Summarising: how can the information be reduced?

Looking at a lot of facts and figures does not always provide you with a clear picture of what is going on. To avoid data overload, it is often a good idea to find a way of summarising the information – perhaps by reducing the many figures to just one representative number.

For example, in many places, especially along a river, one town's waste water discharge may be part of the next town's water supply. It makes sense to monitor the water quality by taking regular measurements of the quality of the river water. This might require hourly measures of the number of milligrams per litre of solids in the water, sampled at many different points on the river. Quite quickly, such a mass of data is generated that it can become difficult to see any underlying patterns. What is needed is some way of reducing many figures into just a few representative ones. Computing a simple daily average, both globally for the entire stretch of the river and locally for each sample point, will provide a useful summary of the levels of pollution.

A second, and equally powerful, way of summarising data is to represent the numbers pictorially using statistical charts or plots – a central theme of Unit 11.

Here are some more examples of investigations of the form 'how many?' or 'how much?':

- How many people die from road accidents each day in the UK?
- What is the typical cost of a tube of toothpaste?
- How old are the students studying MU123?

These are the sorts of questions where a summary in the form of a simple average can really clarify things.

Comparing: is there a difference?

Many of the decisions that we make are based on deciding whether or not there is a difference between two things – does one thing perform better, last longer, or offer better value for money than another?

For example, suppose that in a particular town, some traffic-calming measures are introduced in order to reduce the speed of the vehicles. How would you know whether the initiative was successful? The relevant comparison here is between vehicle speeds before and after the initiative. Let's suppose that a sample of 20 vehicle speeds were recorded both before and after, and that the average speed did indeed fall slightly. What might be your conclusion? One possible explanation might be that the traffic-calming measures have worked. However, there are several problems with this conclusion. First, sample sizes of only 20 are too small to be reliable; one speeding car in the first sample may have made all the difference. Second, it is likely that the speeds of different vehicles vary quite a lot, so differences are to be expected anyway. Third, the difference between the two averages was small. And finally, the lower speeds might have been brought about by some other factor, such as a greater density of

Figure 1 Sampling water quality

Imagine being handed the hourly measurements over a whole month, for twenty different points along the river. For a 30-day month, this would consist of 14 400 numbers!

traffic in the 'after' phase of the experiment, perhaps because it was school term time. So an alternative explanation is that the result is just a matter of chance and that if the experiment were to be repeated, a different conclusion might be drawn. We'll return to this scenario in Activity 3, where you'll be asked to think about how we might perform a more formal statistical investigation into the effects of traffic-calming measures.

In general, investigations involving comparing two averages will depend on several factors, such as the sizes of the samples on which the averages are based, the degree of variation that one might reasonably expect to see in such values, and whether the size of the observed difference is sufficiently large to act upon.

Here are some examples of 'comparing' investigations following on from the previous three examples:

- Do more people, on average, die from road accidents on weekdays or at weekends?
- How does the cost of Brand X toothpaste compare with that of Brand Y?
- Are students studying MU123 older or younger than students on an introductory Arts module?

Seeking a relationship: what sort of relationship is there?

Sometimes a statistical question is not about differences between two or more sets of results but about investigating a possible relationship between quite separate things.

For example, as far back as the 1950s, medical researchers were able to link the number of cigarettes smoked to the incidence of lung cancer. At the time they found that countries with relatively high smoking rates, like the UK, also showed high levels of lung cancer, whereas countries with low smoking levels, like Iceland and Norway, also had low rates of lung cancer. Of course, there were other factors at play that may have affected lung cancer rates, and it is part of the statistician's job to try to isolate and so take account of each of the relevant factors. It is important to remember that a statistical association between two measures does not prove a cause-and-effect relationship between them. For example, the changes in *both* measures may be caused by a *third* factor, such as (in the smoking example) that the citizens of some countries may experience a high level of general stress, which encourages smoking and also contributes to lung cancer.

Sir Richard Doll, one of the scientists famous for establishing the link between smoking and lung cancer, originally suspected tarmac as the cause. Sir Ronald Fisher, an eminent statistician and leading sceptic about the link, suggested a genetic link between lung cancer and the propensity to smoke.

The first paper to propose the link between smoking and lung cancer was R. Doll and A.B. Hill (1950) 'Smoking and carcinoma of the lung. Preliminary report.', *British Medical Journal*, vol. 2, pp. 739–748.

If there appears to be a relationship between two factors, it is often useful to determine what that relationship is. That is, how much does one factor change relative to the other? Here are some more examples in which we might investigate whether there is a relationship between the factors under consideration:

- Are the numbers of road deaths in different countries linked to their respective maximum speed limits?
- How does the cost of tubes of toothpaste depend on their size?
- What is the connection between the numbers of hours that students work on a level 3 module in mathematics and their final grade?

Classifying statistical investigations

Three types of investigation have been described above:

- summarising
- comparing
- seeking a relationship.

Activity 1 asks you to distinguish between summarising investigations and those of other types.

It will be convenient to use the single word 'relationship' to point to the third type of investigation.

Activity 1 *Summarising or otherwise?*

Which of the following investigations are summarising investigations, and which belong to one of the other two categories – that is, either comparing or relationship investigations? Do not try to distinguish between the latter two categories just yet.

(a) How much does a typical loaf of bread cost?

(b) Do men earn more than women?

(c) How heavy is a typical bag of potato crisps?

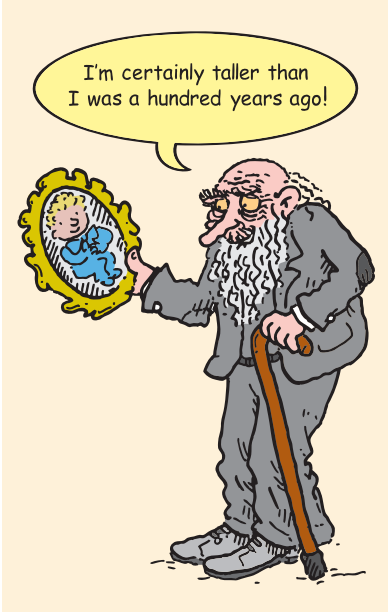
(d) How do grades in an exam depend on the social backgrounds of the students who take the exam?

(e) Is there a link between income level and ill-health?

(f) Did the introduction of car seat-belts save lives?

(g) What proportion of MU123 students are female?

(h) Are people taller than they were 100 years ago?



While summarising investigations are fairly easy to pick out, it can be less easy to distinguish the other two. For example, suppose that an investigation was to be set up to look into the question: 'Do people with long legs tend to run faster than people with short legs?'

Depending on how the investigation was approached, this could be based either on comparing or on seeking a relationship. For example, one possible approach would be to identify two separate groups of people, those with long legs and those with short legs, and compare the running speeds of the two groups. This would be an investigation based on comparing. However, an alternative experimental design could be to choose a sample of people randomly, measure the running speed and leg length of each person, and see if there is a relationship between these two measures. This would be an investigation based on seeking a relationship.

Activity 2 asks you to revisit the five investigations in Activity 1 that were *not* based on summarising and try to classify them into one or other of the two remaining types.

Activity 2 *Comparing or seeking a relationship?*

Classify the investigations below into the 'comparing' and 'relationship' types.

(a) Do men earn more than women?

(b) How do grades in an exam depend on the social backgrounds of the students who take the exam?

(c) Is there a link between income level and ill-health?

(d) Did the introduction of car seat-belts save lives?

(e) Are people taller than they were 100 years ago?

Exploring questions like those above gives a purpose and a direction to statistical learning. In this unit and Unit 11, you will concentrate on questions of the first two types above – summarising and comparing. The third category of question, seeking a relationship between two things, is not explored in great detail in this module, although there is a little on this in Unit 6.

1.2 The statistical investigation cycle

As was mentioned in the Introduction to this unit, there are four clearly identifiable stages in most statistical investigations, which can be summarised as follows.

The four stages of a statistical investigation

Stage 1 **P**ose a question
Stage 2 **C**ollect relevant data
Stage 3 **A**nalyse the data
Stage 4 **I**nterpret the results

It may be helpful to think of these stages set out as a cycle, the PCAI cycle, as illustrated in Figure 2. Here the problem starts in the real world and is resolved by making a journey into the statistical world and back again. Complete resolution of the problem might require several trips around the cycle.

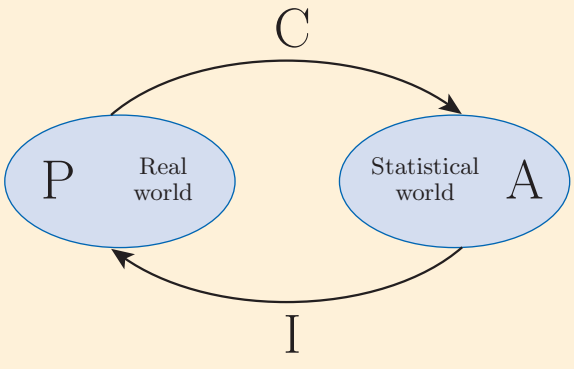


Figure 2 The PCAI statistical investigation cycle

Note that it is important to be as specific as possible in posing a statistical question. The more focused the question, the better the investigation can be attuned to the question, and the better the chance of obtaining a useful answer. The less focused the question, the wider the investigation and the greater the chance that nothing very informative will come out of it.

Not surprisingly, different statistical techniques apply to different stages of the statistical modelling cycle.

- P: Posing a clear question should normally be the first stage of any statistical work. The decision as to which techniques are to be used subsequently will depend on the sort of question that has been asked at the start of the investigation.
- C: Collecting relevant data will involve issues such as choosing samples and designing questionnaires.
- A: This stage, analysing the data, is where techniques like calculating averages and plotting graphs and charts will take place.
- I: The final stage, interpreting the results, takes the action back to the original context from which the initial question was posed. Does the data analysis help to answer the original question? If 'yes', then you can stop there. If 'no', then you may need to travel around the cycle once again, perhaps this time with a slightly modified question or using different analytical techniques.

Activity 3 *Applying the PCAI cycle to a traffic-calming investigation*

Consider a possible investigation, mentioned on page 177, into whether traffic-calming measures are successful in reducing vehicle speed. Spend a few minutes performing a 'back-of-an-envelope' design of this investigation, thinking about the various tasks involved. Then answer the questions below.

That is, jot down a few ideas. It's amazing how many useful things can be (and have been) written, drawn or scrawled on the back of an envelope!

(a) Try to write a description of how this investigation might be conducted, using the stages of the PCAI cycle set out above.

(b) This investigation was introduced as an example of a comparing investigation. What kind of statistical techniques do you think might be involved in the 'A' stage in this case?

The discussion in this subsection should all be rather reminiscent of the four-stage mathematical modelling cycle described in Unit 2; see Table 1 below.

Table 1 A comparison of the four stages of statistical and mathematical modelling

Stage	Statistical modelling	Mathematical modelling (as in Unit 2)
1	P: pose question	Clarify question or problem
2	C: collect data	Make assumptions; collect data
3	A: analyse data	Use mathematics to describe the problem and obtain results
4	I: interpret results	Interpret and check results

Activity 4 asks you to think further about three of the four stages of the PCAI cycle and what sorts of statistical work might be linked to each one.

Activity 4 *Organising the tasks of an investigation*

Here are nine of the common types of task that tend to arise in the C, A and I stages of a statistical investigation. Try to match each task to one of these three stages and then fill in the table below.

- Calculate an average
- Calculate a percentage
- Choose a set of values, or sample
- Make a decision based on an observed, numerical difference
- Design a questionnaire
- Draw a conclusion
- Draw a helpful graph
- Key the data into a spreadsheet
- Make a prediction about the real world

C, A and I stages of a statistical investigation

Collect relevant data:

Analyse the data:

Interpret the results:

Activity 5 *Thinking through the stages of an investigation*

Conventional wisdom suggests that clouds tend to act as a warm blanket, keeping heat in at night and preventing the ground temperature from dropping too far. But is there any evidence, either way, with which to answer the question: 'Do clouds keep heat in?'

Spend up to ten minutes thinking about how you might investigate this question, and then use the four PCAI headings to organise your ideas. Note that you are not asked to carry out this investigation but just to think through the stages involved.

This section concerned two ways of thinking about and addressing statistical questions.

The first was the categorisation of statistical investigations into three types: summarising, comparing or seeking a relationship.

Second, the PCAI statistical investigation cycle was introduced. This cycle, which may be gone around more than once, consists of four stages: posing a question (P), collecting relevant data (C), analysing the data (A), and interpreting the results (I).

2 Dealing with data

This section looks at some issues to do with collecting data (stage 'C' of the PCAI cycle) as well as some important distinctions between different types of data, which have relevance when it comes to stage 'A', analysing the data.

' "Data! Data! Data!", he cried impatiently. "I can't make bricks without clay." '

From *The Adventure of the Copper Beeches*, a Sherlock Holmes story by Sir Arthur Conan Doyle, first published in 1892.

2.1 Primary and secondary data

Having identified the question of interest in the 'P', pose a question, stage of the PCAI cycle, how might you go about the 'C' stage – collect relevant data?

You might consider collecting some data yourself. This might be particularly appropriate if the question of interest is one very specific to you and your surroundings or on which you can collect relevant data quite easily. Data that you collect yourself are called **primary data**. For more substantial research questions, this tends to be a reasonable approach only if 'yourself' refers to a research team in a university, company or other research unit.

As was mentioned at the start of Subsection 1.1, 'Types of statistical question', before rushing into collecting data about any question, it is always a good idea to ask: 'Has anyone collected data on this before?' The answer is often 'yes'! **Secondary data** are data that already exist and can be used or adapted for your purpose.

In this information age, secondary data are plentiful and often readily available through the internet, published literature and other sources. However, inevitably the quality of such data is highly variable. There are a number of consistently reliable sources such as UK government statistics, which are generally professionally collected and presented and free from bias. For example, a useful source of statistical data is the website for the Office for National Statistics (ONS).

However, other sites are set up by organisations that may want to sell you some product or promote a particular set of ideas. In some cases, the data that they present may be subject to bias and distortion, and such sites are best avoided as sources of reliable secondary data.

Data are usually presented as 'datasets'. A **dataset** is a collection of data, usually presented in tabular form, or as a single row, or sometimes as a single column. In this unit you will see all three ways of presenting datasets.

An important convention when presenting any dataset, whether primary or secondary, is to provide an accurate reference to the data source (so that the reader can check the details if they so wish).

Backache in pregnancy

Table 2 contains an extract of data taken from a larger, secondary dataset collected at the London Hospital (now Royal London Hospital). It was designed to help answer questions concerning backache in pregnant women, including: How common is it and how severe? Which factors affect it? Which factors alleviate it?

Table 2 Backache in pregnancy dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Patient number	Back pain severity	Month of pregnancy pain started	Age (years)	Height (m)	Weight at start of pregnancy (kg)	Weight at end of pregnancy (kg)	Weight of baby (kg)	Number of children	Relieved by aspirin?	Relieved by hot bath?	Aggravated by fatigue?	Aggravated by bending?
2	1	1	0	26	1.52	54.5	75	3.35	0	0	0	0	0
3	2	3	0	23	1.6	59.1	68.6	2.22	1	1	0	0	0
4	3	2	6	24	1.57	73.2	82.7	4.15	0	0	0	1	0
5	4	1	8	22	1.52	41.4	47.3	2.81	0	0	0	0	0
6	5	1	0	27	1.6	55.5	60	3.75	1	0	0	0	0
7	6	1	7	32	1.75	70.5	85.5	4.01	2	0	0	0	0
8	7	1	0	24	1.73	76.4	89.1	34	0	0	0	0	0
9	8	1	8	25	1.63	70	85	4.01	1	0	0	1	0
10	9	2	6	20	1.55	52.3	59.5	3.69	1	0	0	0	0
11	10	2	8	18	1.63	83.2	90.9	3.3	99	1	0	0	0
12	11	1	0	21	1.65	64.5	75.5	2.95	0	0	0	0	0
13	12	1	0	26	1.55	49.5	53.6	2.64	0	1	0	1	0
14	13	2	6	35	1.65	70	82.7	3.64	7	0	1	0	1
15	14	1	8	26	1.6	52.3	64.5	4.49	1	0	1	0	0
16	15	1	6	34	1.68	68.2	77.3	3.75	3	0	0	1	0
17	16	0	0	25	1.5	47.3	55	2.73	1	0	0	0	0
18	17	1	7	42	1.52	66.8	73.2	2.44	6	0	0	0	1
19	18	2	6	26	1.65	70	81.4	3.01	1	0	0	0	0
20	19	2	6	18	1.6	56.4	70	3.89	1	0	0	0	0
21	20	0	1	42	1.65	53.6		2.73	2	0	0	0	0
22	21	2	0	28	1.63	59.1	72.3	3.75	99	0	0	0	0
23	22	1	0	26	1.52	44.5	56.4	3.49	0	0	0	0	0
24	23	1	0	23	1.57	55.9	60.9	3.07	0	0	0	0	0
25	24	2	6	21	1.55	57.3	77.3	3.35	0	0	0	99	99
26	25	2	7	32	1.52	69.5	75.5	3.64	5	0	1	0	1
27	26	1	8	18	1.6	73.2	81.4	2.05	0	0	0	0	1
28	27	1	0	25	1.7			2.44	1	1	0	0	0
29	28	1	0	29.916666	1.63	62.7	72.3	3.07	4	0	0	0	0
30	29	1	0	19	1.92	73.6	92.7	3.35	0	0	0	0	0
31	30	2	7	26	1.65	70	89.1	3.21	1	0	0	1	1
32	31	1	8	28	1.68	56.69905	70.9	3.41	0	0	0	0	0
33	32	1	0	21	1.6	58.2	69.5	3.3	0	0	0	0	1
34	33	0	0	29	1.57	68.2	7.5	3.35	0	0	0	0	0

Source: M.J. Mantle et al. (1977) 'Backache in pregnancy', *Rheumatology and Rehabilitation*, no. 16, pp. 95–101, quoted in C. Chatfield (1988) *Problem solving: a statistician's guide*, London, Chapman and Hall.

In order to make the dataset manageable for your work in this unit, the number of respondents has been reduced from 180 women to 33, and the number of items of information reduced from 33 to 13. Note also that the data have been laid out in a spreadsheet format, with numbered rows (1, 2, 3, ...) and lettered columns (A, B, C, ...) which will facilitate identifying particular items of data by their column/row references.

In the remainder of the unit, this dataset will be referred to as the *backache dataset*.

This dataset will be used to illustrate most of the issues concerned with handling data in this section, and to that end, a few of the data values from the original source have been changed. Several of the columns in this table have been entered into the module software resource Dataplotter but, for reasons that will be explained shortly, a few of the data values from this table have been changed. (Also, you won't get around to directly considering the questions concerning backache here as we will focus on making sense of the numbers.)

Take a quick look at these data. The first thing to notice is that each row corresponds to results for one patient, and each column – except the first – to a specific item measured. The first column just contains patient reference numbers. (Notice that because there are column headings in row 1, the patient reference numbers are unfortunately not the same as the table row numbers – a common occurrence.)

Activity 6 *Scanning the data*

Look carefully at the 13 column headings in Table 2 and try to get a sense of what each is measuring. Then try to come up with a few impressions that strike you about the variations in the numbers in the table.

2.2 Discrete and continuous data

The distinction between 'discrete' and 'continuous' measures is important as it provides useful information about the nature of the data collected. As an introduction to these terms, here is a context that should help you get a sense of how they are used.

Note that the word 'discrete' is not the same as 'discreet', which means 'unobtrusive and restrained'.

Look at the picture in Figure 3, which shows a route through a forest. The path itself is continuous, so any position on the path is possible, whereas the stepping-stones placed on the path are discrete; they represent distinct, separate positions with nothing in between any two consecutive steps. Using the path, you might mark your journey in terms of a measured distance, whereas taking the same journey on the stepping-stones involves counting out steps (first, second, third, and so on). In general, this distinction between measuring and counting is a useful way of identifying which measures are discrete and which are continuous.

Figure 3 Discrete stepping stones on a continuous forest path

When it comes to statistical data, the same distinction can be made. An example of discrete data is measurement of shoe size. You can talk about shoe sizes of, say, $7\frac{1}{2}$ or 9, but a shoe size of, say, 8.314 cannot occur in practice, since shoe sizes are restricted to either whole or half sizes. Foot length, on the other hand, has no such restriction – it is something that is measured on a continuous scale of measure and therefore produces continuous data.

Turn back to the data in Table 2. One of the clearest distinctions between the numbers in the columns is that in some columns, the numbers seem to be discrete values while in other columns, the numbers seem to come from a continuous scale. **Discrete data** are data that can take one of a particular set of values; such data typically, though not necessarily, take integer values.

Here are some examples of discrete data:

- the number of days in a week on which one takes exercise
- the number of times a particular website is visited in a day
- the quality of a person's recovery after a serious accident when coded 0 for full recovery, 1 for partial recovery, 2 for failure to recover.

Often, as in the first two examples, discrete data arise from a process of counting, sometimes over a limited range of possible integer outcomes (for example, up to a maximum of 7 days in a week), at other times over an unlimited range (for example, the number of 'hits' on a website, at least in principle!). Sometimes, as in the third example, discrete data arise as a convenient way of coding data whose outcome is really some non-numerical category. A widely-occurring example of this is when there are just two categories such as yes/no, pass/fail or true/false. Such data, coded by two numerical values, are said to be **binary data**. The two values are often taken to be 1 and 0.

The 'bi' in binary means 'two' – as in bicycle (a cycle with two wheels).

Activity 7 *Identifying discrete data*

Examples of columns containing discrete data in the backache dataset are listed and explained below.

Columns	Values available	Explanation
Back pain severity	0, 1, 2, 3	0 = 'nil' 1 = 'nothing worth troubling about' 2 = 'troublesome but not severe' 3 = 'severe'
Relieved by aspirin?	0, 1	0 = no 1 = yes

Note that the 'Relieved by aspirin?' question is a binary measure (with just two outcomes).

Which other columns of data from the backache dataset do you think contain discrete data?

Unlike discrete data, **continuous data** can take all the in-between values on a number scale. In theory, and depending on the context, they may take any numerical value from the set of real numbers, either negative or positive. Alternatively, they may be constrained to be positive (e.g. the length of a particular manufactured item) or they may be limited to a finite interval (e.g. the percentage of active ingredient in a particular compound, which can be anywhere between 0 and 100). In Table 2, the columns not identified in Activity 7, namely 'Height', 'Weight at start of pregnancy', 'Weight at end of pregnancy' and 'Weight of baby', contain continuous data, as perhaps should 'Age'. Notice that all of these columns contain data that take positive values.

Mass and weight

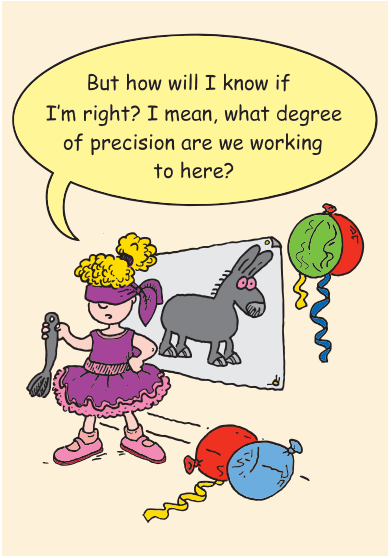
Did you notice that the weights in Table 2 are given in kilograms, even though the kilogram is a unit of mass?

The mass of an object is a measure of the amount of matter that it contains, whereas its weight is a measure of the gravitational force acting on it. Weight, being a force, is measured in newtons. However, you will often see weights quoted in kilograms in everyday life and this informal approach will sometimes be used in MU123 too.

Now, with counts (such as the number of days in a week on which a person takes exercise) and other forms of discrete data, it is possible to give exact answers. With measurements (such as the length of a particular manufactured item), it is never possible to get an exact value, as the next activity illustrates.

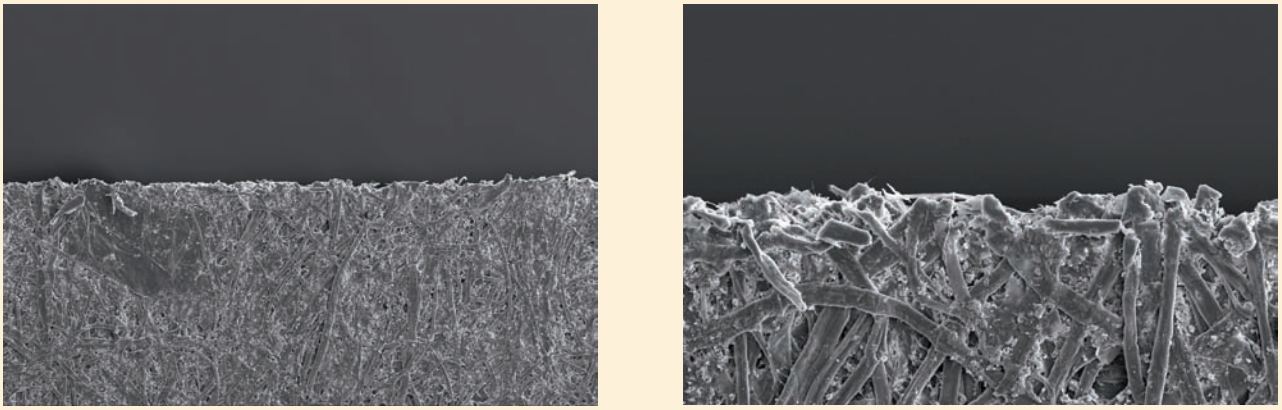
Activity 8 *Investigating measurement precision: how exact is exact?*

Measure the width of a piece of A4 paper with a metric ruler and write down your answer. Please do this before reading any further.



You were deliberately not told in this activity what level of precision to use. You might have written down 21 cm and this would carry the implication that the page width was nearer 21 cm than 20 cm or 22 cm. Alternatively, you might have tried to be more precise and written down 21.1 cm or 211 mm. Again there would be an implication that the actual measurement was nearer 211 mm than 210 mm or 212 mm. If you had access to a more precise measuring device still, you might have been able to write down 211.0 mm or 211.03 mm, and so on.

However, no matter how good your measuring device, you would never be able to say what the *exact* width of the particular sheet of paper was.



(a) (b)

Figure 4 Images showing the ragged edge of a sheet of A4 paper magnified by a factor of (a) 40 and (b) 160

Indeed, as can be seen from Figure 4, at a microscopic level, the edge of a piece of paper is by no means straight and smooth – in fact, the closer we look, the rougher the edge appears to be! Clearly there is a limit to the precision with which is it meaningful to describe its width. However, for most practical purposes, there is no need for extreme precision, and recording the value of the width correct to, say, the nearest centimetre or the nearest millimetre may well suffice.

In practice, for all manufactured items there is a tolerance for the possible range of sizes that each item can be. For example, according to the ISO standard, the width of manufactured A4 paper should be 210 ± 2 mm (i.e. between 208 and 212 mm). This is despite the mathematical exactness suggested in Unit 3 of how ISO paper sizes relate to each other!

This argument about exact measurement applies to other continuous data too: time, age, weight, height, temperature, distance, area, volume or whatever. This is the sense in which each of the columns C to H in Table 2 can be considered to contain continuous data. It is just that the measurement and recording process has resulted in these columns of data being presented correct to the nearest month (when pain started), year (age), centimetre (height), tenth of a kilogram (mothers' weight measurements), and hundredth of a kilogram (babies' weights), respectively.

It follows that while the actual values of two items of continuous data can never strictly be identical, their stated values, given to a certain degree of precision, may well be. For example, patients numbered 15 and 31 in Table 2 are considered to be the same height, 1.68 m or 168 cm, correct to the level of precision measured and recorded.

Activity 9 *Discrete or continuous?*

Which of the following are examples of continuous values and which are discrete?

(a) Price of a loaf of bread, in pence

(b) Number of seagulls on a cliff face

(c) Time of an athlete running 1500 metres, in seconds

(d) Number of goals scored by a hockey team

(e) Distance between major cities, in miles

(f) TMA score achieved by a student

(g) Air temperature at midday at a weather station, in °C

(h) Wind speed measured in kilometres per hour

(i) Wind speed on the Beaufort scale (e.g. gale force 8)

You would be right to think that all measured data are actually discrete, but the idea of continuous data remains useful both conceptually and when creating mathematical and statistical models of the world.

2.3 Checking and cleaning data

In Activity 6, your inspection of the data in Table 2 probably came up with a number of apparent anomalies in the backache dataset. First, there are some blank cells in the spreadsheet where data are missing. Second, you probably also noticed the values of 99 appearing in columns that otherwise contain only small integer values. These must be wrong: those in cells I11 and I22 correspond to women having 99 children from previous pregnancies; those in cells L25 and M25 are given in answer to yes/no questions.

An explanation for these, other than a typing mistake, is that numbers such as 99 are sometimes used as codes that signal 'value missing'. (They can cause trouble: missing data codes might not always be so easy to spot and might even sometimes mistakenly correspond to reasonable actual values.) Good documentation of the data file should make the presence and value of a 'missing data' code clear, but with secondary data this

information can get lost. Large datasets of real data inevitably contain plenty of missing data, and sophisticated statistical methodology has been developed to cope with these gaps.

Take another look at Table 2 on page 184. Are there any further outstandingly large (or small) values in the backache dataset?

Column H contains the weights of the babies, in kilograms. The minimum weight is 2.05 kg, which is not a great deal less than the next smallest weights, 2.22 kg and 2.44 kg. The largest baby's weight, however, appears to be 34 kg (as you might have spotted in Activity 6). This impossible value must surely be the result of a recording error. Most probably, the decimal point was missed out of a weight of 3.4 kg (but without confirmation from the original data collection source, there is no certainty that this is the explanation).

Activity 10 *Examining unusual values in columns of data*

Scan the following columns of the backache data in Table 2 and comment on whether or not you think there might be a problem with any of the most extreme values in each column.

(a) Column G, weights of the mothers at the end of their pregnancies (in kg).

(b) Column E, height (in m).

Outliers

One or more data values that are considerably smaller or larger than the other values in the same dataset are called **outliers**. Sometimes, outliers correspond to errors and it may be possible to correct them and thus remove the outliers. However, as in the case of the tallest mother mentioned in the solution to Activity 10, often there is no such obvious reason and the outlier may just be an unusual, but not unreasonable, observation. You might still wish to ignore or underplay the outlier to come to conclusions about the rest of the data without the outlier influencing results too strongly, or you might wish to embrace the outlier as an important aspect of the data. Either way, again, there are sophisticated statistical techniques available to deal with outliers but these are not explored here.

Figure 5 Spot the outlier!

2.4 Spurious precision

In the backache dataset, you probably noticed the values 29.916666 in cell D29 and 56.69905 in cell F32. These are given with five or six decimal places, whereas other entries in columns D and F are given with zero and one decimal places, respectively. These values are surely examples of **spurious precision**. In the first case, the data value seems to be a result of the person's age being given as 29 years and 11 months, and 11 months being $11/12 = 0.916666$ years (given to six decimal places). It is likely that this value was entered as '$= 29 + 11/12$', which would automatically be displayed in its decimal form. In the second case, the spurious precision has arisen by conversion, to kilograms, of data measured in different units (pounds).

The term 'spurious' means 'different from what it claims to be'.

Activity 11 *Converting spurious precision into appropriate precision*

(a) Given that 1 pound is equivalent to 0.45359237 kilograms, use a calculator to convert a weight of 125 pounds to kilograms. Express the answer correct to five decimal places.

(b) What should the values in cells D29 and F32 be when (appropriately) rounded to zero and one decimal places, respectively?

Spurious precision can arise in various ways. One way, which was illustrated in Activity 11, is in the conversion of units – particularly between metric and imperial measures. For example, a newspaper report may state: 'the flood water was 1 metre (3.2808 feet) deep . . . '. Another way is to imply that a quantity can be measured to a greater level of precision than is possible with the measuring instrument used. For instance, a household ruler may be used to measure lengths to the nearest millimetre, so it would be incorrect to state a measurement to the nearest tenth or hundredth of a millimetre, if the ruler is used.

Another way that spurious precision can arise is when figures are quoted to a greater number of significant figures than is warranted in the context. An example is contained in the following statement which appeared in a newspaper report of a court case in 2005:

> It was estimated that, over a nine-and-a-half year period, the defendant stole £557 327.11.

The accused was a council employee who regularly stole a portion of the money she was counting from the fees paid into machines by motorists in car parks. Think about the quotation for a moment. Do you believe it? Did she really keep a careful record of all the money taken and add it all up accurately? Ah, no, the figure was 'estimated' – but by whom and how?

Activity 12 *Where did the eleven pence come from?*

It is most implausible that the amount of money stolen was exactly £557 327.11.

(a) How do you think this figure might have been arrived at?

(b) In what sense is the figure spurious, and what might be a more reasonable estimate?

The point about using a rounded figure for the amount stolen, £550 000, is that it gives a fairer impression of the degree of precision of the data. Generally speaking it is customary, when analysing data gathered by others, to assume that the claimed precision is justified unless there is definite evidence to the contrary. Furthermore, in displaying data it is frequently unnecessary to retain their full precision. A key principle here is that displayed data should be just precise enough to reveal the key features – offering the reader an answer containing too many significant figures can easily obscure these patterns in a mass of numbers.

If you are collecting primary data, you should bear in mind the kinds of difficulty with data discussed above, and do your best to avoid them. In the case of secondary data, ideally you should be able to go back to the original data collector and check with them any suspicions you have about the data. Unfortunately, all too often this is not possible; with the passage of time, details concerning data collection tend to be forgotten or lost.

2.5 Single and paired data

Column H of the backache dataset in Table 2 contains 33 values of the weights of babies (in kg). This constitutes a single sample of weights of babies from mothers, many of whom suffered from backache in pregnancy. For the moment, ignore the remainder of the data in the table. Looking just at column H, these values are all based on a single measure (weight) and can be described as **single data**. These 33 numbers can be summarised and represented in a variety of ways. You could calculate an average: you will learn more about averages in Section 3, where two different types of average are explored (the mean and the median). Alternatively, you could measure how widely dispersed the values are – in other words, whether the values are tightly clustered together or widely spread. Three measures of spread are explained in Section 4. Finally, you could plot the values to discern the overall pattern visually, and you will be shown a number of useful statistical plots in Unit 11. The purpose of doing these things would be to try to gain an insight into baby weights in general.

Suppose now that there was a second sample, of birth weights from a different set of mothers, the babies in this second sample all being classed as premature. This is now a **two-sample**, as opposed to one-sample, dataset. Interest now is in comparing features of the birth weights of premature babies with those of full-term babies. Other examples of making statistical comparisons might include making a comparison between two medical treatments or two commercial products. Again, a sample of measures would be taken from each and the results compared. (When making such a statistical comparison, there is no requirement that the two samples contain the same number of values, although they could do.)

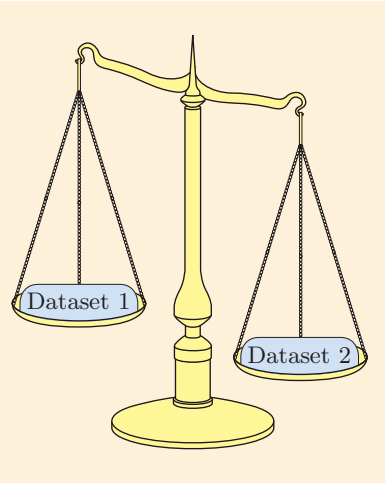


Figure 6 With two-sample data we are interested in comparing

Now return to the babies recorded in the backache dataset in Table 2. A statistical question of interest might be how the babies' weights (in column H) relate to the weights of their mothers at the start of their pregnancy (column F). This question links two pieces of information for each of the people in the study – a case of **paired data**. Interest centres on how one of these pieces of data (birth weight) relates to the other (mother's weight). In statistical investigation terms, this falls under the general heading of seeking a relationship. Although each measure can be explored individually, the main point of collecting paired data is to assess the relationship between the two variables in question.

A characteristic feature of paired data is that the two data lists must contain the same number of values. Also, for each item in one list, there is a specific corresponding item in the other.

A number of important statistical ideas are linked to exploring relationships to do with paired data (for example, regression and scatterplots), and some of these ideas are discussed in Unit 6.

To end this section on properties of data, tackle Activity 13, which asks you to think more deeply about the various data types and how they are typically used to draw sensible conclusions in a statistical investigation.

Activity 13 *Investigating mothers' weights*

Columns F and G of the backache dataset contain the weights at the beginning of pregnancy and at the end of pregnancy for the same set of mothers (ignoring the problem of missing data). Below are four descriptions of possible datasets.

- Weights of mothers and of their babies
- Weights of mothers at end of pregnancy
- Weights of two samples of babies, one in the UK and one in France
- Weights of mothers and average earnings in 20 EU countries

Classify each of these datasets as two-sample data, paired data, single data or two unrelated samples. Also, classify the relevance of each of these for statistical investigation. (One is of no direct interest; one is useful for summarising, one for comparing, and one for seeking a relationship.)

Lay out your solution in tabular form, by filling in the blank spaces below.

Datasets	Data type(s)	Relevance
Weights of mothers and of their babies		
Weights of mothers at end of pregnancy		
Weights of two samples of babies, one in the UK and one in France		
Weights of mothers and average earnings in 20 EU countries		

In this section you looked at dealing with data – in particular, how to classify and distinguish different types of data. Primary and secondary data are terms that identify the data source – primary data you collect yourself, whereas secondary data are taken from somewhere (or someone) else. Data can be discrete, like shoe sizes, or continuous, like the measurement of foot length. Subsection 2.3 explored questions to do with precision, and then introduced the important idea of spurious precision, where values are displayed to a greater-than-warranted number of significant figures. Finally, you looked at single and paired data. Single data are usually represented by a single set of values. Paired data, as the name suggests, involve two related datasets with each data value in one dataset corresponding to a data value in the other dataset. As you will see later in this unit, with single data, the aim is usually to summarise a set of values in order to get a handle on where they lie or to compare two sets of values to see whether one set is bigger or more widely spread than the other. With paired data there is usually a different aim – to investigate a possible relationship between the two measures.

In the next two sections, we will concentrate on single data and how to summarise the information that they contain.

3 Summarising data: location

In this section and the following one, you will take some first steps into stage 'A' of the PCAI cycle, analysing the data that have been collected. In particular, in this section you will look at simple summary measures of what statisticians often call the **location** of a dataset, that is, a single number that represents an 'average', 'typical' or 'central' value.

Activity 14 *Location, location, location*

A group of school students challenged their teachers to a speed-texting competition. The 12 participants (5 teachers and 7 students) each sent a short text message on a mobile phone. The times are given below, and for convenience, the data have been sorted in order of size.

Teacher times (seconds):	18	27	31	36	47		
Student times (seconds):	19	19	21	24	25	27	29

Source: adapted from A. Graham (2006) *Developing thinking in statistics*, London, Paul Chapman Publishing, p. 82.

(a) If you had to summarise each set of times by a single number without doing any calculations, what might you say? (Please bear in mind that there is not a definitively correct answer here.)

(b) Can you make any general comparison of these datasets on the basis of your answer to part (a)?

A statement that describes whereabouts a dataset lies (e.g. 'about 25 seconds') is a way of describing the dataset's location. Activity 14 illustrated the two main purposes of measuring location:

- summarising a set of data values by a single number that might be thought of as an 'average' or 'typical' or 'central' value
- comparing sets of data values on the basis of their locations, to see which set tends to have bigger values.

Measuring language

In the following example, the context is investigating people's use of the language of chance.

At an Open University summer school, a group of 30 students were asked to investigate their understanding of various words used to describe degrees of likelihood (terms such as likely, impossible, nearly certain, fifty-fifty, and so on).

The students were studying the module 'Developing mathematical thinking at Key Stage 3'.

The task began with a consideration of just two such words, 'possible' and 'probable'. Each student was asked to make a personal, numerical estimate on a scale from 0 to 100 (where 0 means impossible and 100 means certain) of their interpretation of these two words as measures of likelihood. The question they were asked to investigate was:

This corresponds to the first stage of the PCAI cycle (discussed in Subsection 1.2): *posing a question.*

> What do people understand by the words 'possible' and 'probable'?

Or, more specifically,

> What numerical values do people attribute to the words 'possible' and 'probable'?

This is an example of a summarising investigation. Before looking at the student data, try this exercise for yourself in Activity 15.

Activity 15 *Investigating words for likelihood*

This task can be thought of as the second stage of the PCAI cycle: *collecting relevant data.*

(a) On a scale from 0 to 100, write down your estimates of the degree of likelihood suggested by the words 'possible' and 'probable'. (You might think of one or both of these words as describing a range of numerical values, but for the purposes of this activity please select a single number reflecting the 'centre' of such a range.)

(b) In your opinion, which of these words would most people rate as describing a higher level of likelihood?

(c) In your opinion, which of these words would generate scores that showed the greater measure of agreement among the respondents?

There are no comments on this activity.

The students' data are contained in the module resource Dataplotter (which you need not open on your computer yet) and shown in Table 3 below. Notice that these are paired data in the sense that each 'pair' corresponds to the response of a particular student. However, for the purposes of this unit, the values will be treated as two-sample data.

Table 3 Thirty students' interpretations of 'Possible' and 'Probable'

Student	Possible	Probable	Student	Possible	Probable
1	30	95	16	10	75
2	90	90	17	50	80
3	60	60	18	20	80
4	70	60	19	50	75
5	5	70	20	50	99
6	1	60	21	30	90
7	1	76	22	1	51
8	50	80	23	30	90
9	50	75	24	40	75
10	50	75	25	30	70
11	80	80	26	20	75
12	30	70	27	50	90
13	1	99	28	30	70
14	10	90	29	98	95
15	85	85	30	35	75

3.1 Scanning data

As you saw in Section 2, before performing a detailed analysis of any dataset, it is always a good idea to scan the data, looking closely at the numbers to see if any patterns or anomalies stand out.

Activity 16 *Scanning Table 3*

(a) Inspect the data in Table 3 and, referring back to the comments on Activity 6 for reminders of the kinds of things you might look for, comment on the presence or otherwise of each of the following:

(i) missing data

(ii) spurious precision

(iii) the constraint that the data lie between 0 and 100

(iv) the presence of outliers.

(b) Does any other feature of the numbers in Table 3 stand out?

Inspecting the data needn't stop there, however. In Activity 17 you are asked to get a first feel for the locations of these sets of data just by further inspection.

Activity 17 *Further inspection of Table 3*

(a) Look closely at the sets of 'Possible' and 'Probable' values in Table 3 and write down what you think would be a typical 'central value' for each one. Then think for a moment about how you came up with this figure – for example, did you do a rough calculation or did you try to pick out a typical value or use some other approach?

(b) Based on your estimates in part (a), which set of values shows the higher location?

3.2 Measuring location

The most important and useful summary of a dataset is a measure of its location, based on some sort of average or typical value. There is no single and universally most appropriate measure of location, but there are various useful measures that can be chosen, depending on the situation and on the nature of the data. Each measure has its own pros and cons. The three most common measures of location in statistics textbooks are the *mean*, the *mode* and the *median*. In this module we will mainly use the mean and the median. These are considered in detail below.

The mode of a dataset is the data value that occurs most frequently. Since there can be several 'most frequent' values, it is a rather problematic measure, which we shall not use in MU123.

You may find the diagram in Figure 7 a useful summary of the ideas set out in the previous paragraph. We will return to this diagram at the end of the next section, by which time a second important summary will have been added, namely the idea of spread and its associated measures: *range*, *interquartile range* and *standard deviation*.

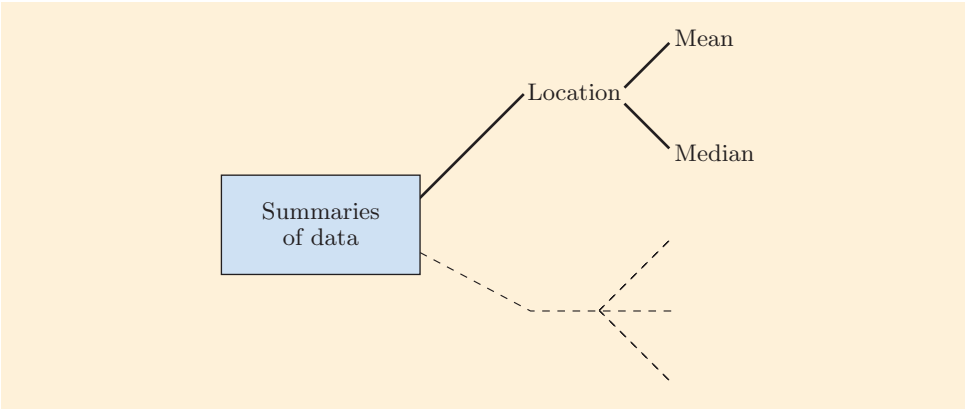


Figure 7 Summaries of data: the mean and median are useful measures of location

The **mean**, which is often just called the average, is probably quite familiar to you.

The mean is more specifically called the **arithmetic mean**, and when applied, as here, to data, it is often called the **sample mean**.

Strategy *To find the mean of a dataset*

To find the mean of a set of numbers, add all the numbers together and divide by however many numbers there are in the set.

Let's look at a simple example, after which you can try a calculation for yourself.

Example 1 *Calculating the mean of a dataset*

Find the mean of the texting times of the five teachers given in Activity 14 (page 193).

Solution

Add together the numbers in the 'Teacher times' dataset and divide by however many numbers are in that dataset.

The mean is

$$(18 + 27 + 31 + 36 + 47)/5 = 159/5 = 31.8 \text{ seconds.}$$

Activity 18 *Finding means*

(a) Calculate the mean of the texting times for the seven students given in Activity 14 (remember to include units in your answer).

(b) Calculate the mean of the 'Possible' values given in Table 3, giving your answer correct to one decimal place.

In Activity 14(a), you may well have taken the approach of selecting the middle value from the five teacher times, which, as you may remember, were listed in increasing order of size. If so, you obtained a statistical average known as the **median**. Speaking roughly, the median is the data value that is in the middle when the data are arranged in order. A more precise definition is given in the following box.

Strategy *To find the median of a dataset*

To find the median of a set of numbers:

- Sort the data into increasing or decreasing order.
- If there is an odd number of data values, the median is the middle value.
- If there is an even number of data values, the median is defined as the mean of the middle two values.

If the middle two values are equal, then the mean is just this common value.

Activity 19 *Finding medians*

(a) Calculate the median of the texting times for the seven students of Activity 14.

(b) How do the medians of the teacher and student texting times compare with the corresponding approximate values of location given in the Comment on Activity 14(a) and the means calculated above?

Activity 20 *More medians*

(a) Twelve sixteen-year-olds were asked to guess the size of the population of the UK and came up with the following estimates, in millions.

60 100 25 60 60 100 80 160 58 23 60 200

Order these numbers and then calculate the median.

(b) Here is an ordered list of the 30 'Possible' values given in Table 3:

1 1 1 1 5 10 10 20 20 30 30 30 30 30 30
35 40 50 50 50 50 50 50 50 60 70 80 85 90 98

What is the median of these values?

Activity 21 *Calculating the 'Probable' mean and median*

This activity, and the next one, represent the third stage of the PCAI cycle: *analysing the data*.

Here is an ordered list of the 30 'Probable' values given in Table 3:

51 60 60 60 70 70 70 70 75 75 75 75 75 75 75
76 80 80 80 80 85 90 90 90 90 90 95 95 99 99

(a) What is the mean of these values?

(b) What is the median of these values?

(c) How do the mean and median compare?

As you are probably already thinking, calculating the mean by 'hand' or calculator becomes increasingly tedious and error-prone as the size of a dataset increases. Calculating the median is a lot easier *if* the data are already sorted. However, this too becomes a long calculation if you have to order a large set of values. So, for anything but the smallest sets of data, it seems appropriate to turn to a calculator, spreadsheet or other software to compute these summary values. Fortunately, these and similar summary calculations can also be easily carried out using the Dataplotter software.

Using Dataplotter to measure location

Using the instructions in Subsection 2.4 of the MU123 Guide, open Dataplotter, which is available on the MU123 website. You can see that four types of plot are available: Dotplot, Boxplot, Histogram and Scatterplot. Ensure that the first of these, Dotplot, is selected. With the 'Datasets' tab selected on the left of the screen, you will see two data lists. From the drop-down menu at the top of the first list, select the dataset '# Possible'. By a similar method, select the dataset '# Probable' from the second list. These two lists contain the data provided in Table 3.

As you can see, as soon as each dataset is selected, Dataplotter processes the information in two ways. First, the values are displayed visually as dotplots, a type of statistical plot where each data value is represented by the position of a dot along the number line below it. Dotplots are a useful way of seeing patterns in data at a glance.

The second main outcome of selecting these datasets is that a set of ten key summary values has been automatically calculated and displayed for each dataset. This is the feature that you are asked to look at now by tackling Activity 22.

Dataplotter

Activity 22 *Summaries of location*

(a) By selecting appropriate values from the two summary lists, complete the table below.

Means and medians of students' values for 'Possible' and 'Probable'

Summary	'Possible' scores	'Probable' scores
Mean		
Median		

(b) Based on the information in the table in part (a), try to come to an initial conclusion about the original question: 'What numerical values do people attribute to the words "possible" and "probable"?' Which of the two datasets had the higher location? Were these results consistent with your previous impression?

The conclusions from using either measure of location in Activity 22 are the same: people tend to think that the word 'probable' indicates a higher degree of likelihood than the word 'possible'. This may not be surprising to you and may be in accordance with what you expected when you thought about the question for Activity 15. But that's not all there is to say about these data. So far you have looked at two summaries that concern measures of *location* from the 'Possible' and 'Probable' data – the calculation of mean and median. In Section 4 you will explore some measures of *spread*.

3.3 Mean versus median

You might be thinking that it's all very well being told how to calculate two different measures of location, but which should you use, and when? Well, it has already been suggested that there is no simple universally applicable answer to this question. But there are a few advantages and disadvantages of one measure compared with the other, and in Activity 23 you are asked to think what some of these might be.

Activity 23 *Mean versus median*

(a) Find the mean and the median of the following datasets.

Dataset A:	3	4	5	6	7	8	9
Dataset B:	3	4	5	6	7	8	99

(b) Which of these two averages seems a more appropriate summary for these datasets?

(c) List the advantages of using each location measure (mean and median).

With this question we are entering the fourth stage of the PCAI cycle: *interpreting the results*.

In practice, both the mean and the median are widely used. They often give similar results but can sometimes differ considerably. Typically, when the values of the dataset are bunched towards one or other end of the range of values, there are larger differences between the mean and the median. For example, with earnings data, it is often the case that the extremely high earnings of a small number of very wealthy individuals will drag up the value of the mean and so may give a rather distorted impression of the location of earnings. For this reason, the median rather than the mean tends to be used for summarising earnings. Where the values are symmetrically spread, there will be little difference between the values of the two summaries, in which case, it will not matter much which one is chosen.

To sum up this section, we have discussed summarising a dataset by measuring its location, that is, a number that might be thought of as an 'average', 'typical' or 'central' value. Two particular measures of location were looked at in detail: the mean and the median. The mean of a set of numbers is found by adding all the numbers together and dividing by however many numbers there are. To find the median, first sort the data in order of size. If there is an odd number of data values, the median is the middle value. If there is an even number of data values, the median is defined as the mean of the middle two values. The mean and median are two measures of location that were used as part of an initial study of the 'probability words' datasets. By comparing them, you were able to give support to the notion that the word 'probable' seems to indicate a higher degree of likelihood than the word 'possible'.

In the next section you will look at another important property of data that can be reduced to a single summary measure. It indicates how closely bunched or spread out the values are and is known as the *spread* of the dataset.

4 Summarising data: spread

In Section 3 you looked at the best known and most widely used summaries of a dataset, which provided information about the *location* of the numbers. These were the mean and median. A second basic property of data is how widely **spread** the values are. For example, here is a set of TMA marks of six students (sorted in increasing order):

42 58 60 68 78 92

As you can see, there is a wide spread of marks here, from a minimum value of 42 to a maximum of 92. In the following TMA, the same students' marks were:

60 65 72 74 75 80

This time the spread of marks is much narrower, ranging from a minimum of 60 to a maximum of 80.

If you go on to more advanced study of statistics, you will find that the measures of spread mentioned here continue to play an important role.

In this section you will look at three of the most common measures of spread – the *range*, *interquartile range* and *standard deviation*.

4.1 Range

As you have just seen, a simple way of estimating spread is to scan along the data to find the smallest (minimum, or 'min') and largest (maximum, or 'max') values. The **range** is the difference between these two values. In other words,

$$\text{range} = \text{max} - \text{min}.$$

Let's try an example.

Example 2 *Calculating the range*

Look at the data below, which give the distances, in kilometres, travelled by eleven students to attend an Open University tutorial.

Distances from home

Distance (km):	12	40	26	4	2	18	66	30	45	12	15

Calculate the range of these data.

Solution

By inspecting the data, the maximum value is 66 km and the minimum value is 2 km.

So the range of these data is

$$\text{max} - \text{min} = 66\,\text{km} - 2\,\text{km} = 64\,\text{km}.$$

Now try Activity 24, which asks you to calculate the range for a different dataset and think about how useful it is as a measure of spread.

Activity 24 *Calculating the range*

Below are the earnings, for a particular week, of 15 staff (including the owner) working in a small business.

Earnings data from a small business

Weekly earnings (£):	280	370	305	285	480	1260	210	340
	280	290	315	325	370	360	280	

(a) Write down the minimum and maximum values in this dataset. Hence calculate the range of weekly earnings.

(b) Why might the range be an unhelpful measure of spread for these particular data?

As you saw in Activity 24, the range is sometimes a rather inadequate measure of spread, particularly where there are one or two extreme outliers in the dataset. In the jargon of statisticians, the range could be referred to as a 'quick-and-dirty' measure of spread – it is quick and easy to calculate, can sometimes give a useful overall impression, but lacks the subtlety and sophistication of other methods. You will now be introduced to two alternative measures of spread that overcome the basic weakness of the range, namely that it is too easily affected by outliers. These are the *interquartile range* and the *standard deviation*.

4.2 Quartiles and the interquartile range

The earnings data in Activity 24 are not very easy to absorb or interpret as presented. Figure 8 shows the same data displayed as a dotplot.

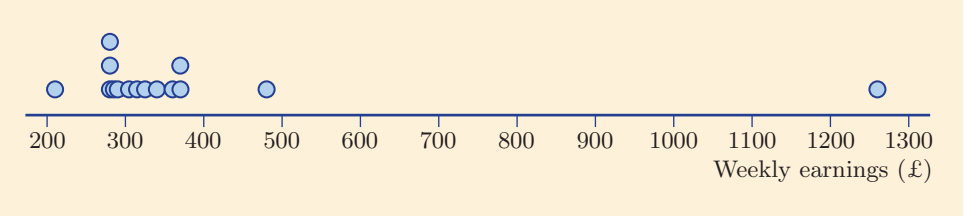


Figure 8 A dotplot showing the earnings data from Activity 24

The 15 dots have been placed above a number line, positioned to correspond to the weekly earnings of the 15 members of staff. Where values coincide (for example, there are three values of £280), the dots are placed vertically, one above another.

When you have a small number of values in a dataset, as is the case here, it is quick and easy to create a simple dotplot of the data like this. Usually it provides a useful, intuitive picture of where the values lie, whether there is some bunching of the data to one side or they are symmetrical, and whether there are clear outliers.

You will use dotplots again in Unit 11.

In this case, the extent to which the outlier is unrepresentative now shows up very clearly.

How then can the problem of the range being unduly affected by this outlier be solved? You might simply decide to ignore this particular untypical value, but that is a somewhat arbitrary decision and not one that can be called a general method, although it is sometimes done.

Alternatively, you might choose to omit, say, the largest and smallest values and take the range of the remaining 13 values. This is a better solution, and one that works well in this particular instance, but if there were several outliers at either end, the problem would not be solved. In order to be confident that you have dealt with the outlier problem, you really need to exclude a greater number of values at either end. The question is, how many?

Introducing the quartiles

The conventional solution, and the one described now, is to exclude the top quarter and bottom quarter of the values and create a new measure of spread that measures the 'range' of the middle 50% of the values. There are two such points that mark the cut-off points of the top and bottom quarters of the data. They are known as the **quartiles** – in particular, the **lower quartile** (Q1) and **upper quartile** (Q3).

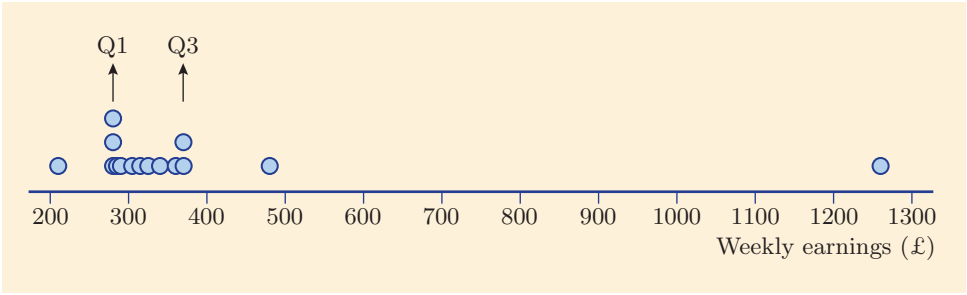


Figure 9 A dotplot of the earnings data from Activity 24 with the quartile positions marked

These quartiles are marked on a dotplot of the earnings data in Figure 9. Take a few moments to satisfy yourself that these values do lie, respectively, roughly one-quarter (Q1) and three-quarters (Q3) of the way through the data.

You will probably find that the description of quartiles is not totally convincing as it rather depends on how we choose to interpret 'a quarter of the way through the dataset'. Incidentally, the median, being the value that lies halfway through the data, is sometimes referred to as Q2, as it is the second quartile.

The convention when defining which quartile is Q1 and which is Q3 is that the data should be presented in increasing order of size. Then, even and odd sample sizes need slightly different approaches, and there are various ways of coping with this. The method for finding the quartiles described in the following two examples is used on some graphical calculators – it is straightforward and quite easy to perform. These examples also show you how to find the measure of spread known as the **interquartile range**, or IQR. The interquartile range is the difference between the upper and lower quartiles, that is, it is the value Q3 – Q1.

Example 3 *Finding the lower and upper quartiles: even sample size*

Find the lower quartile (Q1), the median and the upper quartile (Q3) of the following dataset.

8 3 2 6 4 1 5 7

Then find the interquartile range.

Solution

Sort the data into increasing order. Find the median.

In increasing order, the dataset is

1 2 3 4 5 6 7 8.

The median is the mean of the two middle data values, 4 and 5.

Median = 4.5.

To find the lower quartile, focus on the lower half of the dataset and find the median of this smaller dataset.

The lower half of the dataset is 1 2 3 4.

Its median is 2.5.

So Q1 = 2.5.

To find the upper quartile, focus on the upper half of the dataset and find the median of this smaller dataset.

The upper half of the dataset is 5 6 7 8.

Its median is 6.5.

So Q3 = 6.5.

The interquartile range is the difference between the upper and lower quartiles.

The interquartile range is thus

$6.5 - 2.5 = 4$.

Example 4 *Finding the lower and upper quartiles: odd sample size*

Find the lower quartile (Q1), the median and the upper quartile (Q3) of the following dataset:

1 2 3 4 5 6 7

Then find the interquartile range.

Solution

First find the median.

The median is the middle value of the ordered dataset.

Median = 4.

To find the lower quartile, ignore the middle data value and find the median of the lower 'half' of the dataset.

The lower half of the dataset is 1 2 3.

Its median is 2.

So Q1 = 2.

To find the upper quartile, ignore the middle data value and find the median of the upper 'half' of the dataset.

The upper half of the dataset is 5 6 7.

Its median is 6.

So Q3 = 6.

The interquartile range is the difference between the upper and lower quartiles.

The interquartile range is thus

$$6 - 2 = 4.$$

These examples lead to the following strategy for finding the quartiles and interquartile range.

Strategy *To find the quartiles and the interquartile range of a dataset*

1. Arrange the dataset in increasing order.
2. Next:
   (a) If there is an even number of data values, then the lower quartile (Q1) is the median of the lower half of the dataset, and the upper quartile (Q3) is the median of the upper half of the dataset.
   (b) If there is an odd number of data values, throw out the middle data point (which of course has the median value of the dataset). Then the lower quartile (Q1) is the median of the lower half of the new dataset, and the upper quartile (Q3) is the median of the upper half of the new dataset.
3. The interquartile range (IQR) is Q3 − Q1.

As you have seen, when there is an even number of data values, the dataset breaks neatly in half and the quartiles are simply the medians of these two half-sets. The procedure is slightly more complicated if the original dataset contains an odd number of values, as a decision needs to be made about what constitutes these half-sets. In the strategy above, the data value in the middle is excluded from these half-sets, and this is the convention used on this module. However, the choice of whether or not to include the middle data value is quite arbitrary – some authors include it and others, as we have done here, exclude it. Indeed, there are yet other methods of calculation that are different again and all of these may give slightly different answers for the values of the quartiles. With very small datasets like the ones you have been using, these differences may be noticeable, but in a real investigation, where the sizes would be larger, these small differences tend to disappear.

Activity 25 *Calculating the lower and upper quartiles*

Here again are the earnings data of 15 employees, first introduced in Activity 24. This time, for your convenience, they have been sorted by size.

Weekly earnings (£):	210	280	280	280	285	290	305	315
	325	340	360	370	370	480	1260	

(a) What are the lower and upper quartiles of these values?

(b) Hence find the value of the interquartile range.

4.3 Standard deviation

The best known measure of spread is the **standard deviation**, or SD. The bad news is that, using pencil and paper, it is hard work to calculate the standard deviation, particularly with large datasets. The good news is that, these days, once the data have been keyed in, a calculator or computer can work out the standard deviation in a flash. But before becoming totally reliant on a machine, it is a good idea to perform one or two pencil and paper calculations of the standard deviation using very simple datasets.

An alternative name for the standard deviation is the *RMS deviation* – in full, the **root mean squared deviation**. Literally, it is the (square) root of the mean of the squared deviations. This complicated name will make more sense when you follow through the steps involved in the calculation.

Strategy *To find the standard deviation of a dataset*

1. Find the mean of the dataset.
2. Find the difference of each value from the mean – these are the 'deviations', often labelled as the d values.
3. Square each deviation – this gives the d^2 values.
4. Find the mean of these squared deviations – this number is the 'mean squared deviation', better known as the **variance**.
5. Find the square root of the variance to get the 'root mean squared deviation' – that is, the standard deviation.

There are in fact two different versions of the standard deviation – the method used here involves the mean found in the usual way of adding together all the numbers in a dataset and dividing by n, the size of the dataset. In the other technique for finding the standard deviation, the sum of the squared deviations is divided by $n - 1$ rather than n. These two methods are used in different circumstances, but a discussion of when it is appropriate to use each one is beyond the scope of MU123. In this module, the divisor n will always be used when calculating the standard deviation.

Example 5 *Finding a standard deviation*

Find the standard deviation of the following dataset.

1 2 4 6 7

Solution

Find the mean.

Mean $= (1 + 2 + 4 + 6 + 7)/5 = 20/5 = 4$.

Subtract the mean from each data value to find the deviations.

The deviations are -3, -2, 0, 2, 3.

Square the deviations.

The squared deviations are 9, 4, 0, 4, 9.

Calculate the mean of the squared deviations to find the variance.

The variance is $(9 + 4 + 0 + 4 + 9)/5 = 26/5 = 5.2$.

The standard deviation is the square root of the variance.

So, the standard deviation is

$$\sqrt{5.2} = 2.3 \quad \text{(to 1 d.p.)}.$$

The steps of the calculation may be easier to see when laid out using a table such as the one in Figure 10.

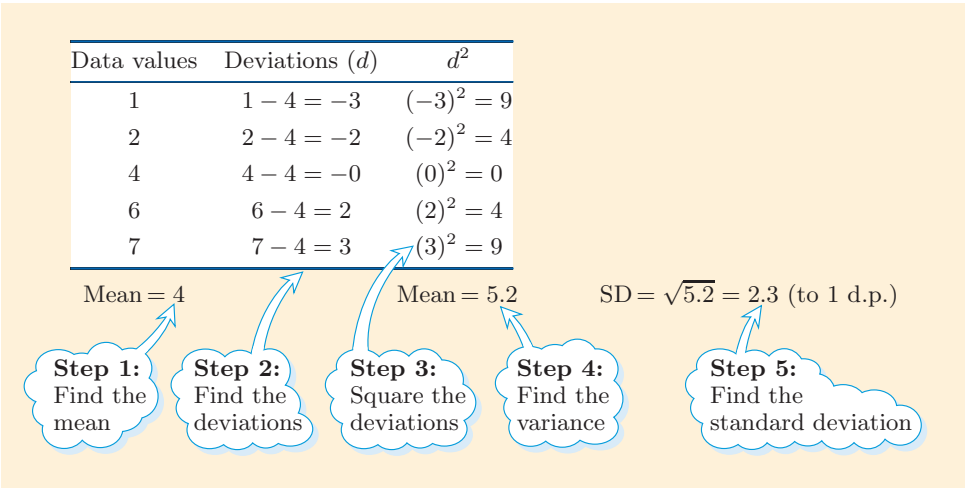


Figure 10 The calculation of standard deviation (SD)

Students often find the calculation of standard deviation rather complicated and the steps hard to remember. It can be helpful to think about some of the ideas visually. Look at Figure 11, which shows these same five data values, 1, 2, 4, 6, 7, in a dotplot. As you can see, the mean, 4, is shown with a vertical line, while the deviations from the mean are represented by horizontal arrows.

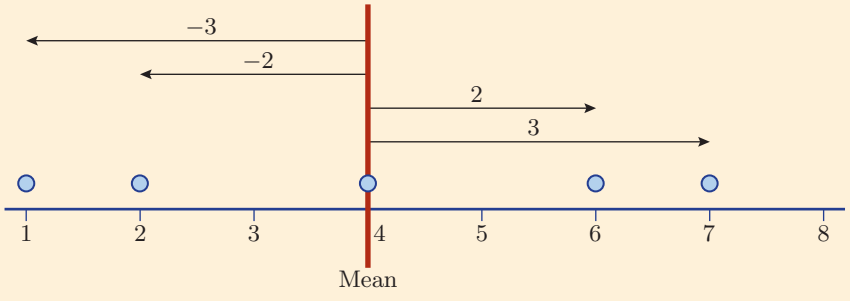


Figure 11 Deviations from the mean

You may have wondered why it is necessary to square the deviations in Step 3 of the calculation. In order to see the point of this, consider what would have happened if you had not squared them. For example, why not just find the mean of these deviations? This would be calculated as follows:

$$\text{mean deviation} = (-3 + (-2) + 0 + 2 + 3)/5 = 0/5 = 0.$$

As you can see, the positive and negative deviations have cancelled each other out and we are left with a numerator of zero. So the value of the mean deviation is zero. In fact, this will always be true of the mean deviation; the positive and negative deviations will always cancel each other out, leaving an answer of zero for the mean deviation. You may like to try this yourself with some other simple numerical examples. For example, take the dataset 3, 4, 6, 11, which has a mean of 6. This produces deviations -3, -2, 0, 5, and again these add to zero.

It is to avoid this problem that the deviations are squared in Step 3 (making them positive), and this is then 'undone' by taking the square root in Step 5.

Now tackle Activity 26, which will give you practice at performing standard deviation calculations using simple datasets.

Activity 26 *Calculating the standard deviation*

Calculate the standard deviations of the following simple datasets.

(a) 1 2 6 11

(b) 2 3 5 6 9

Although calculating standard deviation by pencil and paper is quite hard work, rest assured that these days it is normally done on a calculator or computer; as you will see later, the module resource Dataplotter calculates and displays it and other statistical summaries automatically. There are a number of reasons why the standard deviation is a useful measure of spread, and here are two of the main ones.

Reasons for using the standard deviation as a measure of spread

The standard deviation is the best known and most commonly used measure of spread.

All the values in a dataset are included in its calculation.

(However, unlike the interquartile range, its value can be to some extent distorted by outliers.)

4.4 Investigating spread

Just as Dataplotter provided instant summary measures of location (the mean and the median), it also provides the three summaries of spread introduced in this section: range, interquartile range and standard deviation. To end this section, you are asked to return to the two datasets concerning the words 'possible' and 'probable'. But this time you will explore what these three summaries reveal about the *spread* of these two datasets and how this can be interpreted.

Dataplotter

Activity 27 *Summaries of spread*

Return to Dataplotter and choose the same settings that you used in Activity 22, namely with the Dotplot option checked and the datasets '# Possible' and '# Probable' selected.

(a) By selecting appropriate values from the two summary lists on the screen, complete the table below.

Range, interquartile range and standard deviation of students' values for 'Possible' and 'Probable'

Summary	'Possible' scores	'Probable' scores
Range		
Interquartile range (IQR)		
Standard deviation (SD)		

(b) Based on the information in the table in part (a), which of the two datasets had the wider spread? How would you interpret this?

We should now be able to reach a conclusion about the original question: 'What do people understand by the words "possible" and "probable"?' For reference, Table 4 provides you with a handy list of all of the summaries from Dataplotter associated with the 'Possible' and 'Probable' datasets given earlier.

Table 4 Summaries of the 'Possible' and 'Probable' data

Summary	'Possible' scores	'Probable' scores
Min	1	51
Q1	20	70
Median	32.5	75.5
Q3	50	90
Max	98	99
Mean	38.6	78.5
SD	27.1	11.9
IQR	30	20
Range	97	48
n (size of dataset)	30	30

So, taking the mean and median (found in Activity 22) together with the range, interquartile range and standard deviation (calculated in Activity 27), two conclusions can be drawn. There is evidence that:

- people tend to think that the word 'probable' indicates a higher degree of likelihood than the word 'possible' (a conclusion based on comparing locations)
- there is a greater degree of agreement on the meaning of the word 'probable' than on the meaning of the word 'possible' (a conclusion based on comparing spreads).

As a footnote to this investigation, one of the students who carried out the study commented that the meaning of the word 'possible' rather depends on the tone of voice used when saying it and also on the context. Another

said that there was so much variation in its interpretation that 'possible' really seemed to be a useless word for conveying meaning and should be dropped from the vocabulary!

In Activity 28, the final activity of Section 4, you are invited to enter data directly into Dataplotter to explore the properties of the five summary values that have been introduced in Sections 3 and 4.

Dataplotter

Activity 28 *Investigating small datasets*

Open Dataplotter and clear both datasets by clicking on 'New'.

(a) Enter the four numbers $3, 4, 6, 7$ into the first column (press Enter after each data entry). From the displayed list summaries, you will see that the mean and median of this dataset are both equal to 5.

Now think about entering a fifth number that will raise the overall mean from 5 to 6; what must this number be, and what will be the value of the median of the new dataset? Enter this number to see if you are correct.

See Section 2 of the MU123 Guide for more information on using Dataplotter.

(b) Now change the number you entered to 10, if necessary. With the five numbers $3, 4, 6, 7, 10$ entered in the dataset, the range is 7 (that is, $10 - 3$) and the interquartile range is 5 (that is, $8.5 - 3.5$). Alter one of these five numbers to a different whole number so that the range remains unchanged but the value of the interquartile range increases to 6.

To edit a cell entry, click on the cell, type the new value and press Enter.

(c) Click on 'Clear' in the first column. The title of the dataset is displayed under the drop-down box. Click on the title, type 'SD1' into the box and press Enter. Then enter the six numbers $3, 4, 6, 6, 7, 10$ into the list. The value of the standard deviation is 2.2 (to 1 d.p.).

Next, in the drop-down menu of the second data column, select the dataset 'SD1' that you have just created, click on its title below the drop-down box, change its name to 'SD2' and press Enter. In this second dataset, alter the two 6s by the same amount in opposite directions, that is, by adding some non-zero number to one and subtracting the same number from the other.

Compare the displayed summaries of the two datasets: what effect does this change have on the value of the standard deviation? Can you explain why?

Following on from Section 3, where you looked at summarising a dataset by measuring its location, this section looked at measures of spread. Three particular measures of spread were looked at in detail.

- The *range* of a set of numbers is found by calculating max − min.
- The *interquartile range* (IQR) is the range of the middle half of the data and is calculated as Q3 − Q1 (where Q3 is the upper quartile and Q1 is the lower quartile).
- The *standard deviation* (SD) is the square root of the mean of the squares of the deviations of each data value from the mean. (You were also briefly introduced to a fourth measure of spread, the *variance*, which is the square of the standard deviation.)

These three main measures of spread (range, interquartile range and standard deviation) were used as part of the investigation of the 'Possible' and 'Probable' datasets to demonstrate that there seems to be a greater

measure of agreement on the meaning of the word 'probable' than on the meaning of the word 'possible'.

Figure 12 is a development of the figure introduced in Subsection 3.2. As you can see, it sets out some of the main ideas of this unit, namely that location and spread are the two key forms of data summary, that two measures of location are the mean and median, and that three measures of spread are the range, the interquartile range and the standard deviation.

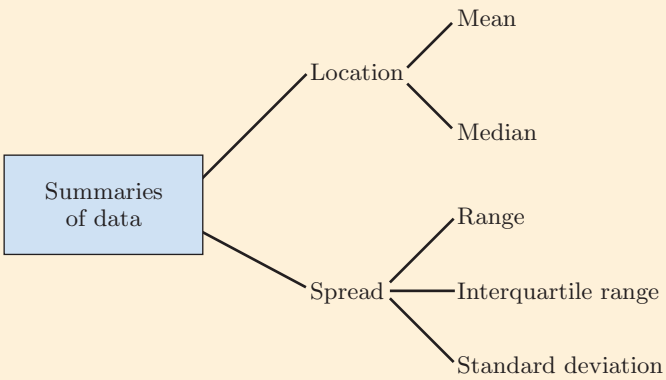


Figure 12 The completed summaries diagram

5 Measuring with accuracy and precision

This unit has looked at some important statistical questions. First, types of statistical questions were categorised (summarising, comparing, seeking a relationship), after which came the idea of a framework for investigating statistical questions (the 'PCAI' framework) – both these notions were explored in Section 1. In Section 2 you saw how to classify and distinguish different types of data (for example, primary and secondary data, discrete and continuous data). Sections 3 and 4 looked at the sorts of measures of location and spread that typically crop up in the 'A' (analyse the data) stage of most statistical investigations. It is these sorts of summary values that really help you to make decisions about data. The second statistical unit of the module, Unit 11, will extend this set of techniques to include a variety of statistical representations in the form of charts and plots.

To end this unit, we look at datasets of a particular kind. It is often useful or necessary to measure the size of a quantity, and this can be done in different ways, depending on the quantity – for example, a length or weight could be measured by using a measuring device, the amount of unemployment in a country or the viewing figures for a television programme could be measured by using surveys, and the strength of gravity could be measured by carrying out an experiment (as you will see shortly). No matter how a measurement is made, it is important to think about how good the measurement is. One way to do that is to consider datasets of *repeated* measurements, and this is the topic of this final section.

First, you are invited to use your own initiative to explore a small dataset created as part of a science investigation. The unit ends with an

examination of two important terms that are often misused and confused: *precision* and *accuracy*, with reference to the statistical summaries *location* and *spread*.

5.1 Summarising a set of scientific measurements

Gravity on the Earth's surface is the downward force exerted on an object by the mass of the Earth. The size of this force varies slightly at different parts of the Earth, in particular at different altitudes and latitudes. When an object is dropped, and if no other forces except gravity are acting on it, its speed increases at a constant rate as it falls to the ground. At sea level, the speed increases by approximately 9.81 m/s every second. This increase in the speed each second is known as the acceleration due to gravity and it is denoted by the letter g. In SI units, g is measured in 'metres per second per second' which is written as $\mathrm{m/s^2}$, so $g \approx 9.81\,\mathrm{m/s^2}$.

The unit $\mathrm{m/s^2}$ can also be written as $\mathrm{m\,s^{-2}}$.

A group of experimenters tried to measure g, based on the following two different methods:

- 'free fall', where a ball bearing is dropped, and the time taken for it to fall through a known height is measured
- 'pendulum', where the period of swing of a pendulum (which is affected by the strength of gravity) is timed.

Posing the question

An interesting question arising out of this experiment is: which of these two methods gives the better estimate for g? You are asked to investigate this, using the data that the experimenters collected.

Collecting the data

The data from the two experiments are given in Table 5 and are also available in Dataplotter. The numbers represent the researchers' results for g, measured in units of $\mathrm{m/s^2}$, based on 16 'free fall' trials and 14 'pendulum' trials.

Table 5 Estimates of g from two experiments

Free fall ($\mathrm{m/s^2}$)	Pendulum ($\mathrm{m/s^2}$)
9.97	10.18
9.84	10.08
9.80	9.78
9.81	9.83
9.80	10.13
9.80	9.95
9.81	9.82
9.81	10.12
9.88	9.96
9.97	9.97
9.78	9.80
9.81	9.81
9.78	9.83
9.80	9.73
9.87	
9.81	

Source: C. Maher and J. Pancari (1990) 'Statistics in high school science', *Teaching Statistics*, vol. 12, pp. 34–7.

At first glance, you may have noticed that there are some subtle variations in the measured values of g shown in Table 5. These variations are not necessarily associated with any geographical differences in g; we can assume that each experiment took place in the same location, so in theory the results should all be the same.

In fact, the differences that we observe in the data can be ascribed to what is known as *experimental error*. All experiments involve a degree of inherent error – inaccuracy in a measurement can arise from a number of sources, for example, poor experimental design, limitations of the measuring equipment, inconsistent application of techniques or even simple human error when reading a measurement. Statistical analysis of repeated measurements, such as calculating the mean of a dataset of repeated trials, is an important method for minimising the effects of experimental error in scientific experiments.

Activity 29 *Scanning the data*

Run your eye down both columns of figures. What general impressions do you have of these figures and what clues do they give about the success of the two experiments?

Dataplotter

Activity 30 *Analysing the data*

Return to Dataplotter. Using the drop-down menus at the top of each list, select dataset '# Free fall' for the first list and '# Pendulum' for the second list.

Use suitable measures of location and spread to decide which of these two experiments produced a better estimate for g. It will suffice to consider all summary measures rounded to two decimal places.

Analysing the data

As was done in Sections 3 and 4 with the 'Possible' and 'Probable' datasets, a variety of measures of location and spread are instantly calculated and displayed in Dataplotter. Table 6 gives the various location and spread summary measures (rounded to 2 decimal places) for these two datasets.

Table 6 Summaries of the 'Free fall' and 'Pendulum' datasets

	Free fall	Pendulum
Min	9.78	9.73
Q1	9.80	9.81
Median	9.81	9.89
Q3	9.86	10.08
Max	9.97	10.18
Mean	9.83	9.93
SD	0.06	0.14
IQR	0.06	0.27
Range	0.19	0.45
n (size of dataset)	16	14

Two features stand out from these summaries.

In terms of location, the averages (i.e. mean and median) of the 'free fall' data lie closer to the 'true' value for g of 9.81 m/s^2 than do the 'pendulum' averages.

In terms of spread, there is a much narrower spread for the 'free fall' than for the 'pendulum' data as calculated by any of the measures.

Interpreting the results

So, on the evidence of these summaries, the 'free fall' experiment produced

better results, since the average of the experimental values was closer to the true value and the results were more closely clustered together.

5.2 Accuracy and precision

A word often used in this unit is 'precision'. Elsewhere in the module it is used in the context of a number being stated to so many decimal places or a certain number of significant figures. Its meaning will be extended here. It is a term that can easily be confused with 'accuracy', but in fact these two terms have a subtle difference in meaning.

Imagine that you have been asked to audition for Robin Hood's 'merry men' and you have passed all the necessary merriment tests. The final set of tests requires you to demonstrate prowess with the bow and arrow. On your first run, with the sun glinting through the old oak tree, your five arrows land as shown in Figure 13.

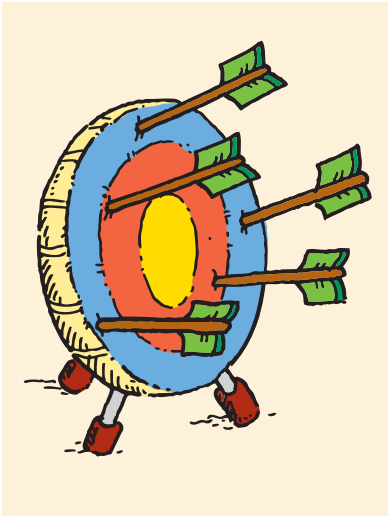


Figure 13 First run: accurate but not precise

How would you describe this performance? The answer is that this shows *accuracy* but not precision. You can be described as accurate, because the average of the five shots is close to the centre. You are not precise, because the shots are widely spread.

Robin wasn't too impressed with your first effort, but he agrees to give you five more shots. This time, there is greater consistency, as shown in Figure 14.

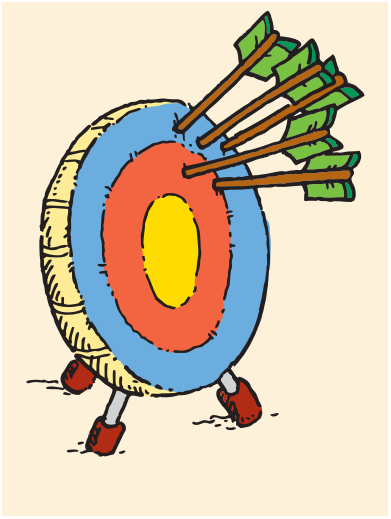


Figure 14 Second run: precise but not accurate

'More consistent, yes,' says Robin, who was known throughout Sherwood Forest for his wit and repartee, 'but you are consistently missing!'

What Robin meant to say was that your shooting shows greater *precision* (the shots now cluster together) but the accuracy is actually worse than before (they are off-centre).

You beg for one last chance – and this time you really show that you have got to grips with the accuracy and precision issues in your archery skills, as shown in Figure 15. The tight clustering on this final run shows that you have lost none of the precision of Run 2, while the centring on the bull's-eye shows that your accuracy from Run 1 has returned.

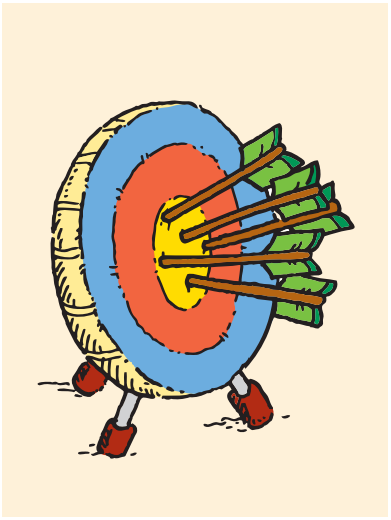


Figure 15 Third run: both accurate and precise

Welcome to Sherwood Forest!

The first moral of this little tale is that accuracy is a statement about the location of a set of measurements: the closer the *average* of the measurements to the true value of what is being measured, the more *accurate* your estimation of that value. The second message is that precision tells you about the spread of that set of measurements: the more tightly *packed* your values, the more *precise* your measurements are.

Accuracy versus precision

For a set of (repeated) measurements:

- **Accuracy** describes how close the average is to the true value.
- **Precision** describes how close the measurements are to each other.

Ideally, when making measurements, you would like to have both accuracy and precision! In Subsection 5.1, the measurements from the 'free fall' experiment were both more accurate and more precise than the measurements from the 'pendulum' experiment.

Activity 31 *Distinguishing between precision and accuracy*

Two kitchen weighing scales are being tested to see if they measure accurately. A 100 g test weight is weighed five times on each set of scales and the results are shown below.

Scales A:	102 g	101 g	102 g	100 g	100 g
Scales B:	98 g	100 g	99 g	99 g	103 g

(a) Calculate the mean and range of each dataset.

(b) Use the measures in (a) to decide which set of scales is more accurate and which is more precise.

This final short section began by asking you to apply your skills in calculating summaries (both of location and spread) to a scientific investigation for estimating the value of g, the acceleration due to gravity. The final subsection looked at two words, accuracy and precision. It suggested that accuracy is a statement about the location of a set of measures, whereas precision tells you about their spread.

To review the ideas in this unit, have a go at the practice quiz for Unit 4 and then try the iCMA and TMA questions.

Learning checklist

After studying this unit, you should be able to:

- distinguish between different types of data, such as: primary and secondary data; discrete and continuous data; single and paired data
- use the PCAI cycle and appreciate how the various statistical skills and techniques fit into the PCAI stages of a statistical investigation
- gain a rough-and-ready, but sensible, overview of a dataset by just scanning the data (inspecting the values by eye)
- check and, if justified, clean data where there are numerical discrepancies such as outliers
- summarise a dataset in terms of measures of location and spread, select suitable data summaries for the context and use them as evidence to draw a conclusion
- appreciate the differences between accuracy, precision and spurious precision and use appropriate rounding in numerical summaries.

Solutions and comments on Activities

Activity 1

Investigations (a), (c) and (g) are in the summarising category. All the remaining investigations are in either the comparing or the relationship category.

Activity 2

The cases (a), (d) and (e) are 'comparing' investigations. Case (c) is a 'relationship' investigation. In case (b), the social backgrounds can be measured on a numerical scale and a relationship investigation carried out. However, the social backgrounds can also be used to split the students into two groups so that a comparing investigation can be carried out.

Activity 3

(a) Although this is a hypothetical example, the potential stages should be fairly easy to identify. Here is one possible answer (yours may be different in several respects).

Stage P: The general question here is: 'Did the traffic-calming measures slow the traffic?' A more focused question might be: 'Were vehicle speeds slower, on average, after the measures were introduced?'

Stage C: A suitable sample of vehicle speeds would be collected before and after the traffic-calming measures were introduced. Care would need to be taken to ensure that the sample sizes were sufficiently large, that the vehicles were chosen randomly, and that the circumstances of the sampling (e.g. time of day) were similar between the two samples.

Stage A: A simple technique for analysing these figures is to calculate and compare the two average speeds. (You haven't seen these yet, but useful techniques at the 'A' stage could include drawing a variety of different charts and plots: for example, dotplots, boxplots and histograms, all of which are covered in Unit 11.)

Stage I: The trickiest stage is relating the data analysis to the original problem. In simple terms, if the average speed is lower after traffic-calming measures are in place, then we might conclude that they have been successful. However, although average traffic speeds may have fallen slightly, how do we know that this was *because* of the traffic-calming measures? There might be some other explanation for these results, such as that the 'after' measurements were conducted in very different weather conditions when traffic conditions could be expected to be slower anyway. Also, the difference may not be sufficiently large to allow us to draw any firm conclusions. It is worth pointing out that although more focused questions are usually easier to answer, they may contain certain in-built assumptions that compromise the original question that you asked. So, as has already been indicated, it is one thing to show that the average speeds after the traffic-calming measures were introduced had fallen, but it is another to prove cause-and-effect.

Furthermore, it is conceivable (though perhaps unlikely) that average speeds are indeed reduced but that this makes no difference to the number and severity of accidents. (Reducing accidents and their severity was probably the purpose of the exercise in the first place.)

(b) Calculating averages and plotting data graphically are very useful statistical techniques, particularly for investigations of comparing.

Activity 4

Collect relevant data:
- Choose a sample
- Design a questionnaire
- Key the data into a spreadsheet

Analyse the data:
- Calculate an average
- Calculate a percentage
- Draw a helpful graph

Interpret the results:
- Make a decision based on a difference
- Draw a conclusion
- Make a prediction about the real world

Activity 5

Clearly there are no uniquely correct answers to this activity, but you might like to compare your notes with the suggestions below.

Stage P: A suitable question might be: 'Do clouds keep heat in?' A more focused question might be something like: 'Do night temperatures drop less when it is cloudy?'

Stage C: Various newspapers publish daily lists of the highest and lowest temperatures recorded on the previous day across a number of towns in the UK, and also whether the weather had been sunny, cloudy, raining, and so on. For a particular town or region, collect this information for 30 consecutive days, noting whether each day was cloudy or clear.

Stage A: Separate the 30 days' data into two groups, labelled cloudy and clear. For each of the 30 days, calculate the 'temperature swing', that is, the difference between the highest and lowest temperatures. Calculate the average temperature swing for the cloudy data and the clear data.

Stage I: Observe which average temperature swing was greater. If the answer was the 'clear' data, this would provide evidence to suggest that clouds do tend to retain heat. However, a weakness with the design of this investigation is that we are not strictly comparing like with like. Since sunny days are generally warmer than cloudy days, an alternative explanation might be that temperature swings are larger on warmer days (as opposed to sunny days, which are not necessarily the same thing).

Activity 6

Probably the most striking features are the following.

- Some columns (such as E to H) seem to contain 'detailed' decimal numbers taking a wide range of values, while all the others seem to contain many fewer integers, some (such as J to M) containing a lot of zeros.
- Some cells (G21, F28, G28) that might be expected to contain data are blank.
- Some individual numbers stand out as being much larger than the rest of the numbers in the same column. For example, the values of 99 in cells I11, I22, L25 and M25 are enormous compared with other numbers in those columns, as is (rather less obviously) the value in cell H8 which corresponds to a baby weighing 34 kg!
- Some individual numbers stand out as being given in a rather different form from the rest of the numbers in the same column. The values in cells D29 and F32 are given to many more decimal places than the other values in their columns.

Activity 7

Discrete data can be found in the columns 'Month of pregnancy pain started', 'Number of children', 'Relieved by hot bath?', 'Aggravated by fatigue?' and 'Aggravated by bending?'. The last three of these contain binary data (except for the values of '99'). You may have added 'Age' to the list – the column contains (mostly) integer values between 18 and 42 – but you may have had a nagging doubt that surely age is measured on a continuous scale; more on this very soon!

Activity 8

Comments on this activity are included in the text.

Activity 9

Discrete:

(a) Price of loaf

(b) Number of seagulls

(d) Number of goals scored

(f) TMA score

(i) Wind speed on the Beaufort scale

Continuous:

(c) Athlete's time

(e) Distance between cities

(g) Air temperature

(h) Wind speed measured in kilometres per hour

In cases (c), (e), (g) and (h) the underlying measures (time, distance, temperature and wind speed) are continuous, but in practice values are measured and recorded on what is effectively a discrete scale.

Activity 10

(a) Column G has a maximum weight of 92.7 kg, which is both a realistic value and in line with the next largest weights, and a minimum of 7.5 kg ... which is not realistic and wholly out of line with other mothers' weights! A possible explanation here is that a decimal point has been erroneously introduced into a true value of 75 kg.

(b) The data in cells E2 to E34 range from 1.5 m to 1.92 m. The latter is very much greater than any other height in this column. However, is it necessarily in error? Such a woman would be unusually tall, but such a height is by no means impossible.

Activity 11

(a) 125 pounds, in kg, is $125 \times 0.45359237 = 56.699046\ldots$ kg, which is 56.69905 kg when given correct to five decimal places, as in cell F32.

(b) 29.916666 years becomes 30 years, and 56.69905 kg becomes 56.7 kg.

Activity 12

(a) The likelihood is that this estimate involved 'scaling up' daily or weekly data. For example, suppose on a particular day the accused was found to have stolen, say, £266.65. Let us assume that the employee worked roughly 220 days per year (44 weeks at 5 days per week). Over 9.5 years, the number of days worked would be $9.5 \times 220 = 2090$

days. Based on the assumption that £266.65 was a typical day's 'takings', a rough estimate of the total theft could be made by calculating £266.65 × 2090 = £557 298.50.

(b) One wonders where the spare 11p or, more particularly, the final 1p would have come from. This is the sort of spurious precision that makes one smell a statistical rat! It seems unlikely that car park pay-and-display machines were accepting pennies in 1996. It would be much more reasonable to accept, say, £550 000 as a rough estimate of the amount of money stolen.

Activity 13

Datasets	Data type(s)	Relevance
Weights of mothers and weights of their babies	Paired data	Seeking a relationship
Weights of mothers at end of pregnancy	Single data	Summarising
Weights of two samples of babies, one in the UK and one in France	Two-sample data	Comparing
Weights of mothers and average earnings in 20 EU countries	Two unrelated samples	No direct interest

Activity 14

(a) You might say that the teacher times were all about 30 seconds, while the student times were all about 25 seconds.

(b) On the basis of the answer to part (a), student times were generally faster than teacher times.

Activity 16

(a) **(i)** There are no missing data.

(ii) There do not appear to be any numbers given to spurious levels of precision in the sense of too many decimal places.

(iii) All the data values lie between 0 and 100.

(iv) You might or might not think of labelling some values as outliers (e.g. the single-figure values for 'Possible' stand out ... but there are several of them). All told, the data seem to be pretty 'clean'.

(b) An interesting observation of human behaviour is that, when asked to make an estimate of something, most people have a tendency to round their answers to, say, the nearest 5 or 10. There seems to be considerable evidence of such a tendency here since a large number of values in both columns are divisible by 5 or 10. (You might particularly have noticed the many numbers ending with a zero. These represent respondents who have applied an appropriate degree of precision to the question asked.)

Activity 17

(a) You might have tried to identify the values that tended to crop up most often, or maybe disregarded the very large and very small values and identified a value that lies in the middle of the remaining items.

(b) Based on inspecting the data and perhaps your own response to Activity 16(b), you may have thought that the 'Probable' values were a bit higher than the 'Possible' values.

Activity 18

(a) The mean texting time for the seven students is

$$(19 + 19 + 21 + 24 + 25 + 27 + 29)/7$$
$$= 164/7 = 23.4 \text{ seconds (to 1 d.p.)}.$$

(b) The mean of the 'Possible' values in Table 3 is

$$(30 + 90 + 60 + \cdots + 35)/30$$
$$= 1157/30 = 38.6 \text{ (to 1 d.p.)}.$$

Activity 19

(a) Where there are seven values sorted in order of size, the median is the fourth value. So the median of the seven student texting times is 24 seconds.

(b) The table below shows the three sets of summaries already used for these data. (The 'Estimate' column refers to the estimated values in the solution to Activity 14(a).)

	Estimate	Mean	Median
Teacher times (s)	30	31.8	31
Student times (s)	25	23.4	24

As you can see, all three teacher averages are fairly similar, as are the three student averages.

Activity 20

(a) In increasing order, the data become:

23 25 58 60 60 60
60 80 100 100 160 200

The median is the mean of the 6th and 7th values in the ordered list, i.e. $(60 + 60)/2 = 60$. As these estimates are millions, the median estimate is therefore 60 million.

Alternatively, in decreasing order, the data values (in millions) are:

200 160 100 100 80 60
60 60 60 58 25 23

The two middle values are still 60 and 60, so again the median estimate is 60 million.

(b) The median is the mean of the 15th and 16th values in the ordered list, namely $(30 + 35)/2 = 32.5$.

Activity 21

(a) The mean is

$(51 + 60 + 60 + \cdots + 99)/30 = 78.5$.

(b) The median is the mean of the 15th and 16th values in the ordered dataset, namely $(75 + 76)/2 = 75.5$.

(c) The mean is larger than the median for this dataset.

Activity 22

(a) *Means and medians of students' values for 'Possible' and 'Probable'*

Summary	'Possible' scores	'Probable' scores
Mean	38.6	78.5
Median	32.5	75.5

(These values have been rounded to 1 d.p.)

(b) There are two pairs of summary values here, each of which gives a direct answer to the question posed.

Means: the 'Possible' mean of 38.6 is well below the 'Probable' mean of 78.5.

Medians: the 'Possible' median of 32.5 is well below the 'Probable' median of 75.5.

Further comments on this activity can be found in the text following the activity.

Activity 23

(a) For Dataset A, mean = 6 and median = 6.

For Dataset B, mean = 18.9 (to 1 d.p.) and median = 6.

(b) Dataset A is perfectly symmetrical – i.e. the values are not bunched together on one side or the other but are located in a way that is evenly balanced around the middle value of 6. Where datasets are highly symmetrical, the values of the mean and the median are very similar, so it really doesn't matter which you choose.

With Dataset B, the outlier 99 has a big impact on the value of the mean, but has no effect on the value of the median. There is no easy answer to which is the better choice of summary value in this case, as it all depends on the context from which the numbers were taken. If you feel that the 99 is a freak value and should effectively be disregarded, then choose the median. However, if the 99 is important and needs to be recognised in the summary, then choose the mean.

(c) Several advantages of the mean and the median are listed below, in no particular order. Please note that what seem to be advantages to some people might seem to be disadvantages to others! Also, you are not expected to have thought of all the pros and cons listed here.

Possible advantages of the mean:

- The mean, or average, is familiar to most people and widely used.
- There is often a button on a simple scientific calculator for calculating the mean, but not one for calculating the median.
- The mean includes every value in its calculation. (This, in particular, may or may not be an advantage!)

Possible advantages of the median:

- When there is an odd number of values in the dataset, the median is one of the values from the dataset and so can be thought of as being a 'representative' of the complete dataset.
- Following on from the point above, you might think that the median is a more intuitive summary (see also the way Activity 14 was approached).
- With reference to small datasets, the median is easier to calculate in your head than is the mean.

- As you saw in part (a) of this activity, an important property of the median is that it isn't affected by outliers: even if, say, the largest value in a dataset is made very much larger than the other values in the dataset, the median, being the 'middle value', doesn't change.
- You can identify the median item even in situations where there are no actual figures. For example, if you want to choose a soldier of average height, simply ask all the soldiers under consideration to line up in order of size and choose the one in the middle.

Activity 24

(a) min = £210, max = £1260, so
range = £1260 − £210 = £1050.

(b) The max value (£1260) is clearly very far out of line with the rest of the data. (It is likely that this figure represents the weekly earnings of the owner.) In fact, most of the values are bunched between £200 and £400, so for this particular dataset, the range does not give a useful impression of the spread of the main part of the data.

Activity 25

(a) The median is the 8th value when placed in order, i.e. £315.

Q1 is the median of the bottom half of the data (excluding the median data value), i.e. the median of

210 280 280 280 285 290 305

So Q1 = £280.

Q3 is the median of the upper half of the data (excluding the median data value), i.e. the median of

325 340 360 370 370 480 1260

So Q3 = £370.

(b) Interquartile range = £370 − £280 = £90.

Activity 26

(a) Mean $= (1 + 2 + 6 + 11)/4 = 20/4 = 5$.

The deviations are found by subtracting the mean from each data value in turn, giving $-4, -3, 1, 6$.

The squared deviations are $16, 9, 1, 36$.

The variance is the mean of these squared deviations:

$$(16 + 9 + 1 + 36)/4 = 62/4 = 15.5.$$

The standard deviation is the square root of the variance:

$$\sqrt{15.5} = 3.9 \quad \text{(to 1 d.p.)}.$$

(b) Mean $= (2 + 3 + 5 + 6 + 9)/5 = 5$.

The deviations are found by subtracting the mean from each data value in turn, giving $-3, -2, 0, 1, 4$.

The squared deviations are $9, 4, 0, 1, 16$.

The variance is the mean of these squared deviations:

$$(9 + 4 + 0 + 1 + 16)/5 = 30/5 = 6.$$

The standard deviation is the square root of the variance:

$$\sqrt{6} = 2.4 \quad \text{(to 1 d.p.)}.$$

Activity 27

(a) *Range, IQR and SD of students' values for 'Possible' and 'Probable'*

Summary	'Possible' scores	'Probable' scores
Range	97	48
IQR	30	20
SD	27.1	11.9

(b) There are three pairs of summary values here, each of which gives a direct answer to the question posed.

Ranges: the 'Possible' range of 97 is much wider than the 'Probable' range of 48.

IQRs: the 'Possible' interquartile range of 30 is much wider than the 'Probable' interquartile range of 20.

SDs: the 'Possible' standard deviation of 27.1 is much wider than the 'Probable' standard deviation of 11.9.

So, in general, the spread of estimates for the 'Possible' data is considerably wider than that for the 'Probable' data. What this suggests is that, if this sample is typical, when people use the word 'possible', it is difficult to know what sort of level of likelihood they are referring to since numerical estimates for defining this word are so widely spread.

Activity 28

(a) The fifth number is 10. This will also raise the value of the median from 5 to 6.

(b) Change the 7 to 9. The min and max values are unchanged, so the range remains at 7. The upper quartile, Q3, increases from 8.5 to 9.5, which increases the interquartile range to 6 (i.e. $9.5 - 3.5$).

(c) Changing the 6s as described should have the effect of increasing the value of the standard deviation. The reason is that 6 happens to be equal to the value of the sample mean, so the two data items 6 each have a deviation of zero. Changing the 6s as described to any other values will produce non-zero deviations for these data values (without changing the mean), which will increase the value of the standard deviation.

Activity 29

At first glance, you can see that there don't seem to be any outliers or examples of spurious precision. A closer look might suggest that the 'free fall' values seem to be slightly lower and less widely spread than the 'pendulum' data.

Activity 30

Comments on this activity are included in the text.

Activity 31

(a) For the Scales A, the mean (in g) is $(102 + 101 + 102 + 100 + 100)/5 = 505/5 = 101$.

The range is $102\,\text{g} - 100\,\text{g} = 2\,\text{g}$.

For the Scales B, the mean (in g) is $(98 + 100 + 99 + 99 + 103)/5 = 499/5 = 99.8$.

The range is $103\,\text{g} - 98\,\text{g} = 5\,\text{g}$.

(b) The mean weight from Scales B (99.8 g) is closer to the true weight of 100g than the mean weight from Scales A (101 g), so the Scales B are more accurate.

The range of the weights from Scales A (2 g) is smaller than the range of the weights from Scales B (5 g), so the Scales A are more precise.